Modélisation et inférence par une approche couplant filtrage et algorithmes stochastiques pour des dynamiques épidémiques partiellement observées

Romain Narci

université PARIS-SACLAY

Modélisation et inférence par une approche couplant filtrage et algorithmes stochastiques pour des dynamiques épidémiques partiellement observées

Modeling and inference by an approach coupling filtering and stochastic algorithms for partially observed epidemic dynamics

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 574, mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques aux interfaces
Graduate School : Mathématiques. Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche MaIAGE (Université Paris-Saclay, INRAE), sous la direction de Catherine LARÉDO, directrice de recherche émérite, la co-direction de Maud DELATTRE, chargée de recherche, le co-encadrement de Elisabeta VERGU, directrice de recherche

Thèse soutenue à Paris-Saclay, le 17 février 2022, par

Romain NARCI

Composition du jury

Chi Viet TRAN Professeur, Université Gustave Eiffel	Président
Pierre-Yves BOËLLE Professeur des universités - praticien hospitalier, Sorbonne Université	Rapporteur & Examinateur
Nathalie PEYRARD Directrice de recherche, INRAE Toulouse	Rapportrice & Examinatrice
Pierre BARBILLON Professeur, AgroParisTech	Examinateur
Mélanie PRAGUE Chargée de recherche, INRIA Bordeaux	Examinatrice
Catherine LARÉDO Directrice de recherche émérite, INRAE Jouy-en-Josas	Directrice de thèse

THESE DE DOCTORAT
NNT : 2022UPASM002

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre : Modélisation et inférence par une approche couplant filtrage et algorithmes stochastiques pour des dynamiques épidémiques partiellement observées

Mots clés : Modèles à effets mixtes, modèles à espace d'état, estimation paramétrique, algorithme SAEM, filtre de Kalman, processus Gaussiens à petite variance

Résumé : Le cadre de cette thèse est l'inférence pour des dynamiques épidémiques partiellement observées. Le développement d'approches pour l'estimation de paramètres de ces dynamiques à partir des observations de leur suivi est un premier enjeu important, cependant difficile du point de vue statistique. En effet, les données sont généralement recueillies à des temps discrets et sujettes à des erreurs de report et/ou de mesure. A cela s'ajoute la présence de composantes non-observées dans les modèles mécanistes servant à décrire les épidémies, rendant nécessaire l'utilisation d'algorithmes pour l'inférence (cadre des modèles à variables latentes). Par ailleurs, des dynamiques multiples d'une même épidémie peuvent être observées dans des sites géographiques distincts ou à des périodes différentes. La prise en compte explicite de la variabilité inter-épidémies dans la modélisation constitue un deuxième enjeu.

Nous considérons des dynamiques épidémiques en population de taille finie modélisées par des processus Markoviens de sauts dépendant de la densité. A travers une approche s'appuyant sur leur approximation par des processus de diffusion, les dynamiques épidémiques sont décrites par des processus Gaussiens à petite variance. En considérant également une approximation Gaussienne du modèle des observations, nous nous plaçons dans le cadre des modèles à espace d'état Gaussiens et linéaires en les états du système. Dans ces modèles, la vraisemblance des observations peut s'obtenir via des techniques de filtrage de Kalman, que nous combinons avec une procédure d'optimisation pour estimer les paramètres. Puis, dans le cas d'épidémies multiples, nous utilisons le cadre des modèles à effets mixtes permettant de décrire de façon plus fine et originale les différentes sources de variabilité entre plusieurs jeux de données recueillis à partir d'épidémies multisites ou récurrentes. Dans ces modèles, des distributions sont spécifiées pour les paramètres, ce qui permet de prendre en compte la variabilité inter-épidémies. Pour estimer les paramètres de ces distributions, nous combinons des techniques de filtrage de Kalman et l'algorithme stochastique SAEM. Dans un premier temps, les différentes méthodes développées dans cette thèse sont évaluées sur des jeux de données simulées d'épidémies décrites par des processus Markoviens de sauts. La mise en œuvre sur des données réelles porte sur l'incidence quotidienne de syndromes grippaux entre 1990 et 2017 en France (épidémies récurrentes) et les données hospitalières relatives à la Covid-19 dans 12 régions de la France métropolitaine pendant le printemps 2020 (épidémies multisites).

Title : Modeling and inference by an approach coupling filtering and stochastic algorithms for partially observed epidemic dynamics

Keywords : Mixed effects models, state space models, parametric estimation, SAEM algorithm, Kalman filter, Gaussian processes with small variance

Abstract : The context of this thesis is the inference for partially observed epidemic dynamics. The development of approaches for estimating parameters of these dynamics from observations of their follow-up is a first important challenge, but difficult from a statistical point of view. Indeed, the data are generally collected at discrete time points and subject to reporting and/or measurement errors. Additionally, the presence of unobserved components in the mechanistic models used to describe the epidemics requires the use of inference algorithms (latent variable model framework). Furthermore, multiple dynamics of the same epidemic event can be observed in different geographic locations or at different times. A second challenge is to explicitly take into account inter-epidemic variability in the modeling.

We consider epidemic dynamics in fixed population size modeled by Markovian density-dependent jump processes. Through an approach based on their approximation by diffusion processes, the epidemic dynamics are described by Gaussian processes with small variance. Considering a Gaussian approximation of the observational model, we adopt the framework of the Gaussian state space models and linear in the system states. In these models, the likelihood of the observations can be obtained via Kalman filtering techniques that we combine with an optimization procedure to estimate the parameters. Then, in the case of multiple epidemics, we use the framework of the mixed effects models allowing to describe in a more refined and original way the different sources of variability between several data sets collected from multisite or recurrent epidemics. In these models, distributions are specified for the parameters, which allows to take into account the inter-epidemic variability. To estimate the parameters of these distributions, we combine Kalman filtering techniques and the stochastic SAEM algorithm. The different methods developed in this thesis are first assessed on simulated epidemic data sets described by Markovian jump processes. The implementation on real data concerns the daily incidence of influenza-like illness between 1990 and 2017 in France (recurrent epidemics) and hospital data related to Covid-19 in 12 regions of metropolitan France during spring 2020 (multisite epidemics).


MATH
INNOV


Région
îledeFrance


MaiAGE
Mathématiques et Informatique Appliquées du Génome à l'Environnement


Fondation mathématique
FMJH
Jacques Hadamard


INRAE

Remerciements

Je tiens tout d'abord à remercier mes encadrantes de thèse Catherine Larédo, Maud Delattre et Elisabeta Vergu à la fois au point de vue scientifique mais aussi humain. L'association de leurs domaines d'expertise riches et variés m'ont permis de vivre une expérience extrêmement enrichissante. De plus, je tiens aussi à les remercier pour leur soutien sans faille tout le long de la thèse (et même avant et après !), et ce, tout particulièrement pendant les moments difficiles. Pour avoir toujours été présentes, je vous remercie vivement.

Je remercie Pierre Yves-Boëlle et Nathalie Peyrard d'avoir accepté de rapporter ma thèse et pour leur bienveillance dans l'écriture de leur rapport. Je remercie aussi Pierre Barbillon, Mélanie Prague et Chi Viet Tran d'avoir participé à ma soutenance ainsi que pour leurs questions auxquelles j'ai pris beaucoup de plaisir à répondre.

Je remercie Chistine Kéribin et Chistophe Giraud qui m'ont guidé lors des années de Master et qui m'ont fait connaître l'unité MaIAGE. Je remercie aussi Mahendra Mariadassou et Stéphane Robin de m'avoir fait confiance pour un CDD d'un an à AgroParisTech avant le début de ma thèse.

Je remercie mes collègues de l'INRAE, dont certains sont devenus mes amis : Ajmal, pour ton sérieux et ta bienveillance envers ton prochain, Léo, pour ta personnalité très attentive, soucieuse et responsable dont j'ai rarement pu observer une telle affirmation, Lina, pour ton calme imperturbable et ta profonde maturité et Henri, pour ton optimisme et ton dynamisme sans faille. Merci à mes ex-collègues doctorants Marie, Marion et Madeleine pour les pauses cafés très agréables de ces derniers mois. Merci à Gildas, Estelle, Olivier, Laurent, Patrick, Béatrice et tant d'autres pour avoir contribué à l'ambiance saine, sereine et très agréable que j'espère retrouver un jour dans le monde professionnel. Je tiens spécifiquement à remercier Laurent pour avoir pris le temps de m'indiquer comment utiliser la machine à reliure !

Je termine avec des remerciements plus personnels. Je remercie mon grand-père de m'avoir fait connaître l'INRAE (anciennement INRA) alors que j'étais encore enfant. Je remercie ma mère, mon père et mon frère dont le lien spécial nous permet de vivre de façon vivante les expériences des autres, et ainsi d'alléger le fardeau pouvant être parfois lourd à porter tout seul. Enfin, je remercie Somaya, mon âme soeur, avec qui je partage la même trame : aucun passé, présent ou futur n'existe sans toi, pour un dénouement final à la hauteur de notre aspiration.

Table des matières

1 Introduction **5**

1.1 Modélisation mécaniste des épidémies . . . 8

1.1.1 Modèles déterministes de propagation d'épidémies . . . 8

1.1.1.1 Modèles compartimentaux classiques . . . 9

1.1.1.2 Modèles avec structure en classes homogènes . . . 10

1.1.1.3 Modèles à temporalité forcée . . . 12

1.1.1.4 Modélisation de mesures de contrôle . . . 12

1.1.1.5 Modélisation avec perte d'immunité . . . 13

1.1.1.6 Modélisation d'épidémies multisites . . . 14

1.1.2 Modèles stochastiques de propagation d'épidémies . . . 15

1.1.2.1 Processus Markoviens de sauts . . . 15

1.1.2.2 Processus Markoviens de sauts densité-dépendants . . . 17

1.1.2.3 Processus de diffusion en grande population . . . 18

1.1.2.4 Approximation Gaussienne . . . 19

1.1.2.5 Conséquences statistiques . . . 19

1.2 Modélisation des observations . . . 20

1.2.1 Modélisation des données de prévalence et d'incidence . . . 21

1.2.1.1 Prévalence . . . 21

1.2.1.2 Incidence . . . 21

1.2.2 Modélisation du bruit d'observation . . . 23

1.2.3 Modélisation simultanée de plusieurs dynamiques épidémiques . . . 25

1.2.3.1 Contexte . . . 25

1.2.3.2 Modèles à effets fixes . . . 27

1.2.3.3 Modèles à effets aléatoires ou mixtes . . . 27

1.3 Quelques algorithmes pour l'inférence . . . 29

1.3.1 Algorithmes pour les modèles à espace d'état . . . 30

1.3.1.1 Filtre de Kalman . . . 30

1.3.1.2 Méthodes particulaires . . . 31

1.3.1.3 Algorithme MIF . . . 32

1.3.2 Algorithme Expectation-Maximisation et ses variantes . . . 33

1.3.2.1 Algorithme EM . . . 33

1.3.2.2 Algorithme MCEM . . . 34

1.3.2.3 Approximation stochastique de l'EM . . . 35

1.4 Objectifs et contributions de la thèse . . . 38

2 Premier article : Inference for partially observed epidemic dynamics guided by Kalman filtering techniques. **42**

2.1 Introduction . . . 43

2.2 Gaussian model approximation for large population epidemics . . . 45

2.2.1 Preliminary comments on inference in epidemic models . . . 45
2.2.2 Approximation of large population epidemic models and the autoregressive point of view . . . 45
2.2.3 Approximation of the observation model . . . 48
2.2.4 Application on the SIR epidemic model . . . 49
2.3 Parameter estimation using Kalman filtering techniques . . . 51
2.3.1 Approximate likelihood inference . . . 51
2.3.1.1 Preliminary results in the general framework of Kalman filtering 52
2.3.1.2 Recursive computation of the approximate log-likelihood . . . 52
2.3.2 Application on the SIR epidemic model . . . 53
2.4 Simulation study . . . 53
2.4.1 Simulation settings . . . 53
2.4.2 Inference : settings, performance comparison, and implementation . . . 54
2.4.3 Point estimates and standard deviations for key model parameters θ . . . 56
2.4.3.1 Simulation results for the first experiment ($\tau = 0$) . . . 56
2.4.3.2 Simulation results for the second experiment ($\tau \neq 0$) . . . 57
2.4.3.3 Additional comments . . . 60
2.4.4 Confidence interval estimates based on profile likelihood . . . 61
2.5 Application on real data . . . 62
2.6 Discussion . . . 63

3 Deuxième article : Inference in Gaussian state-space models with mixed effects for multiple epidemic dynamics 66
3.1 Introduction . . . 67
3.2 A mixed-effects approach for a state-space epidemic model for multiple epidemics 69
3.2.1 The basics of the modeling framework for the case of a single epidemic . 69
3.2.2 Modeling framework for multiple epidemics . . . 72
3.3 Parametric inference . . . 73
3.3.1 Maximum likelihood estimation . . . 73
3.3.2 Convergence of the SAEM-MCMC algorithm . . . 76
3.4 Assessment of parameter estimators performances on simulated data . . . 76
3.4.1 Simulation setting . . . 76
3.4.2 Point estimates and standard deviations for inferred parameters . . . 78
3.4.3 Comparison with an empirical two-step approach . . . 80
3.5 Case study : influenza outbreaks in France . . . 81
3.6 Discussion . . . 85

4 Estimation dans des modèles à effets mixtes : application sur des données épidémiques régionales de la Covid-19 en France 88
4.1 Introduction . . . 89
4.1.1 Contexte : pandémie de Covid-19 . . . 89
4.1.2 Bref état de l'art des modèles dynamiques Covid-19 en France . . . 89
4.1.3 Objectifs et démarche . . . 90
4.1.4 Description des données . . . 91
4.2 Modélisation compartimentale de propagation de la Covid-19 . . . 91
4.2.1 Modèle mécaniste (formalisme EDO) . . . 92
4.2.2 Expression du R_0 à partir des paramètres du modèle . . . 93
4.3 Inférence . . . 94
4.3.1 Modèle pour une épidémie . . . 94
4.3.1.1 Modélisation statistique des variables d'état . . . 95

4.3.1.2 Modélisation statistique des observations . . . 95
4.3.1.3 Stratégie d'estimation des paramètres du modèle . . . 96
4.3.2 Modèles à effets mixtes . . . 97
4.3.2.1 Variabilité intra-épidémie . . . 97
4.3.2.2 Variabilité inter-épidémies . . . 98
4.4 Etude empirique de l'identifiabilité du modèle . . . 98
4.4.1 Design des simulations . . . 98
4.4.2 Simulations hiéararchiques . . . 98
4.4.3 Réglages algorithmiques . . . 99
4.4.4 Résultats . . . 99
4.5 Application à l'épidémie de la Covid-19 dans 12 régions de la France métropolitaine 99
4.6 Discussion . . . 102

5 Conclusion et perspectives **103**

6 Annexes **107**
A Annexes : Chapitre 2 . . . 108
A.1 Remarks on the sampling interval . . . 108
A.2 Proof of Proposition 2 . . . 108
A.3 Proof of Lemma 2 . . . 109
A.4 Proof of Proposition 3 . . . 109
A.5 Additional simulation study . . . 109
A.5.1 Description . . . 109
A.5.2 Point estimates and standard deviations for key model parameters θ . . . 110
A.5.3 Numerical confidence intervals . . . 115
A.6 User-friendly code . . . 116
B Annexes : Chapitre 3 . . . 117
B.1 Key quantities involved in the SEIR epidemic model . . . 117
B.2 Details on the Kalman filter equations for incidence data of epidemic dynamics . . . 117
B.3 Practical considerations on implementation setting . . . 118
B.4 Estimation results for a second set of parameter values . . . 119
B.4.1 Simulation settings . . . 119
B.4.2 Point estimates and standard deviation for inferred parameters . 120
C Annexes : Chapitre 4 . . . 123
C.1 Modèles mécanistes utilisés dans la littérature pour décrire la propagation du SARS-CoV-2 dans une population en France . . . 123
C.2 Tableau résumé de plusieurs études d'inférence . . . 126
C.3 Visualisation de l'évaluation post-prédictive pour $\tau_A = 0.4$ et $\tau_A = 0.6$ . . . 128

Chapitre 1

Introduction

Table des matières

1.1 Modélisation mécaniste des épidémies 8
1.1.1 Modèles déterministes de propagation d'épidémies 8
1.1.2 Modèles stochastiques de propagation d'épidémies 15
1.2 Modélisation des observations 20
1.2.1 Modélisation des données de prévalence et d'incidence 21
1.2.2 Modélisation du bruit d'observation 23
1.2.3 Modélisation simultanée de plusieurs dynamiques épidémiques 25
1.3 Quelques algorithmes pour l'inférence 29
1.3.1 Algorithmes pour les modèles à espace d'état 30
1.3.2 Algorithme Expectation-Maximisation et ses variantes 33
1.4 Objectifs et contributions de la thèse 38

Préambule

L'étude de la propagation des maladies infectieuses correspond à des enjeux importants du point de vue sanitaire et économique, l'émergence d'épidémies pouvant entraîner des coûts sociétaux et humains considérables (cf Heesterbeek et al. (2015)). A titre d'exemple, la pandémie du virus SARS-CoV-2 a eu des conséquences dramatiques sur de nombreux aspects de la société (santé, économie, culture, éducation, etc.) avec plus de 200 millions de cas confirmés dans le monde dont plus de 4 millions de décès (WHO (Octobre 2021)). La grippe humaine, infection respiratoire contagieuse provoquée par des virus influenza, possède un mécanisme saisonnier et constitue aussi un problème majeur de santé publique. En France, elle touche chaque année 2 à 8 millions de personnes et provoque entre 10,000 et 15,000 décès (Pasteur (2020)). En 2014, la maladie à virus Ebola a provoqué en Afrique de l'Ouest une épidémie d'ampleur considérable, avec au moins 28,000 cas officiellement déclarés, dont plus de 11,000 décès (Pasteur (2021)). Chez les animaux d'élevages, les maladies infectieuses constituent aussi un problème de santé animale et économique, et parfois même de santé publique (e.g. maladies à prions). Les maladies infectieuses endémiques des bovins, telles que la paraturberculose bovine, se transmettent principalement par des transferts d'animaux entre fermes, entraînant d'importantes pertes animales et économiques (Carslake et al. (2010)). Dans un tel contexte, il est important d'élaborer des outils permettant de mieux comprendre les mécanismes sous-jacents à la propagation d'épidémies afin de guider les politiques de santé publique et vétérinaire (Keeling and Rohani (2007)).

Une approche naturelle pour décrire la propagation d'une maladie infectieuse consiste à élaborer une modélisation mécaniste de ses différentes dynamiques, comportant des paramètres les caractérisant. Par exemple, un indicateur très souvent utilisé est le *ratio de reproduction de base*, R_0, qui mesure le nombre moyen de cas secondaires générés par un individu infectieux typique dans une population entièrement saine. Si $R_0 > 1$, l'épidémie a une probabilité non nulle de croître exponentiellement tandis que si $R_0 \leq 1$, l'épidémie ne se développe pas de façon majeure et s'éteint rapidement. Ce paramètre peut varier en fonction du *taux de transmission* λ, défini comme le produit entre le taux de contact entre individus et la probabilité de transmission lors d'un contact. Un autre paramètre important concerne la *période moyenne d'infectiosité*, c'est-à-dire l'intervalle de temps durant lequel un individu est contagieux, définie comme l'inverse du *taux de guérison* γ (Keeling and Rohani (2007)). Les paramètres de ces modèles mécanistes ont des valeurs généralement inconnues. Par conséquent, un enjeu statistique consiste à proposer des approches permettant d'estimer ces paramètres clés afin de mieux caractériser les processus épidémiques et, à terme, fournir des prédictions fiables de leurs dynamiques. Ici, la difficulté est d'adapter ces approches à la nature des observations disponibles.

En effet, deux types de données épidémiques existent :

— données de *prévalence* : nombre (ou proportion) total(e) de cas à plusieurs temps d'observation ;

— données d'*incidence* : nombre (ou proportion) de *nouveaux* cas survenus entre deux temps consécutifs.

Ces données sont récoltées par des organismes gouvernementaux, des plateformes de surveillance ou des réseaux collaboratifs associés à des instituts de recherche. Par exemple, le réseau Sentinelles (url : `http://www.sentiweb.fr`), système d'information basé sur un réseau de médecins généralistes bénévoles en France métropolitaine, surveille un certain nombre de maladies infectieuses telles que les syndromes grippaux (surveillance remplacée depuis 2020 par celle des infections respiratoires aiguës), la diarrhée aiguë ou la varicelle. De plus, en lien avec ces processus de collecte des données de suivi épidémique, les données disponibles comportent de nombreuses spécificités

qu'il est important de prendre en compte dans la modélisation statistique des épidémies. Dans cette thèse, nous nous concentrons sur trois spécificités des observations :

(i) la *partialité des données* : seuls certains états de santé sont observés (individus infectés et contagieux, hospitalisés, etc.) tandis que d'autres ne sont pas ou sont difficilement mesurables pour diverses raisons (individus infectés mais asymptomatiques, etc.) ; données recueillies de façon journalière ou hebdomadaire et souvent agrégées temporellement et/ou spatialement ;

(ii) les *erreurs d'observation* : les données épidémiques sont sujettes à des erreurs d'observation telles que le sous-report du nombre de cas et les erreurs de diagnostic ;

(iii) *les variabilités inter- et intra-épidémie* : les données épidémiques peuvent être variables en fonction de l'endroit ou de la période où elles sont récoltées (e.g. les techniques de report peuvent évoluer au cours du temps ou selon la région considérée ; les caractéristiques épidémiques telles que le taux de transmission ou la période d'infectiosité peuvent être variables dans le temps dû à l'évolution de l'agent pathogène ou à la prise de mesures de contrôle).

Pour détailler le cadre statistique approprié décrivant les dynamiques de propagation d'infections et le processus d'observation, nous utilisons le formalisme des *modèles à variables latentes* dont un sous-ensemble est constitué par les *modèles à espace d'état* (e.g. Cappé et al. (2005)). Dans ces modèles, $(X_k)_{k\in\mathbb{N}}$ est un processus de Markov *caché*, c'est-à-dire qui n'est pas observable, dont les transitions dépendent d'un vecteur de paramètres η. Conditionnellement à la chaîne de Markov (X_k), $(Y_k)_{k\geq 1}$ est une séquence de variables aléatoires indépendantes observées dont la distribution (conditionnelle) dépend d'un vecteur des paramètres μ. Dans ce cas, la suite (X_k) et le couple (X_k, Y_k) sont Markoviens, mais la suite (Y_k) ne l'est pas. La Figure 1.1 schématise la structure de dépendance de ces modèles. L'objectif est alors d'estimer η et μ à partir des observations disponibles (Y_k).

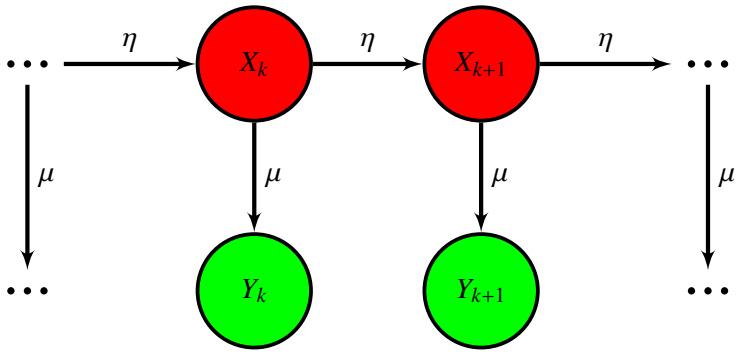

FIGURE 1.1 – Représentation graphique d'un modèle à espace d'état générique, où (X_k) est la variable cachée (en rouge), (Y_k) est la variable observée (en vert), η est un vecteur de paramètres mécanistes et μ est un vecteur de paramètres d'observation.

Les chapitres de ce manuscrit s'articulent autour de la modélisation de dynamiques épidémiques et de l'inférence des paramètres du modèle. Dans la suite de l'introduction, nous décrivons les ingrédients habituels utilisés en modélisation des dynamiques épidémiques avant de proposer un état de l'art partiel des méthodes d'inférence utilisées en épidémiologie ou dans le cadre plus général des modèles à variables latentes (cf Section 1.3). Dans les Sections 1.1 et 1.2, deux types de modèles sont présentés : les modèles mécanistes des épidémies (correspondant à la version continue du

modèle d'état (X_k)) et de leurs observations (décrivant la distribution de Y_k conditionnellement à X_k). Le premier type est utilisé pour décrire les mécanismes de transmission d'agents pathogènes tandis que l'intérêt du deuxième est d'offrir un cadre statistique pour l'inférence de paramètres. La Section 1.4 présente les objectifs et contributions de la thèse.

1.1 Modélisation mécaniste des épidémies

Il existe de nombreux modèles mathématiques pour décrire des dynamiques épidémiques de transmission de pathogènes, qui ont pour objectif principal d'accroître la connaissance existante sur la propagation des maladies et informer sur les effets d'éventuelles mesures de contrôle (Keeling and Rohani (2007)). Ces modèles sont de complexités variables, les plus simples reposant sur des hypothèses simplificatrices de la réalité tandis que les plus complexes peuvent inclure de nombreux éléments spécifiques tels que des cycles sociaux (vacances, Cauchemez et al. (2008)), des interventions des pouvoirs publiques ou des variations climatiques (e.g. dengue : Cazelles et al. (2005), choléra : Ionides et al. (2006), grippe humaine : Shaman and Kohn (2009)). La difficulté consiste ici à trouver un compromis entre d'une part, la complexité du modèle et d'autre part, sa capacité à répondre aux objectifs recherchés (e.g. estimer des paramètres épidémiques clés, fournir des prédictions fiables) et sa flexibilité d'utilisation dans des conditions réelles. A partir du livre de Keeling and Rohani (2007) et de l'article de revue de O'Neill (2010), nous présentons plusieurs modèles mécanistes existant dans un formalisme déterministe ou stochastique.

Le processus de modélisation de la propagation d'une maladie dans une population d'individus peut comporter plusieurs couches de complexité, choisies spécifiquement selon l'agent pathogène étudié, son histoire naturelle et celle de la maladie et l'objectif scientifique recherché. Dans les modèles à compartiments, offrant un cadre structurel pour l'étude des épidémies (Kermack et al. (1927)), chaque compartiment dénombre l'effectif (ou la proportion) d'individus dans un certain état de santé vis-à-vis d'un pathogène donné en fonction du temps. L'évolution de ces effectifs dans le temps peut être formalisée par un système dynamique déterministe ou stochastique. Dans ce qui suit, par souci de clarté, ces modèles et leurs extensions sont détaillés sous le formalisme déterministe. Le formalisme stochastique est quant à lui présenté dans un cadre plus générique.

1.1.1 Modèles déterministes de propagation d'épidémies

Les systèmes d'équations différentielles ordinaires (EDO) constituent un formalisme naturel pour décrire les modèles à compartiments dans une population fermée de taille N. L'hypothèse implicite est que les dynamiques de transmission sont déterministes, c'est-à-dire que pour un jeu de paramètres et de conditions initiales donnés, les dynamiques résultantes des modèles sont identiques. Nous commençons par décrire les modèles les plus simples afin d'en donner ensuite des extensions incorporant d'autres états de santé et/ou des couches de complexité additionnelles. En assimilant à $X(t)$ le vecteur des d états du modèle compartimental au temps t, les systèmes d'EDO sont définis sous la forme générique suivante :

$$\frac{dx}{dt}(t) = b(x(t)), \quad x_0 \neq (0, \dots, 0)^t, \tag{1.1}$$

où $x(t) = \frac{X(t)}{N}$ correspond aux proportions d'individus dans chaque compartiment, x^t est la transposée de x, $x_0 = \frac{X_0}{N}$ et $b(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ est une fonction des états du système. Dans la suite, quand il n'y a pas d'ambiguïté, nous notons $X := X(t)$, $x := x(t)$ et $b(x) := b(x(t))$ par souci de clarté.

Remarque : Lorsque la solution de (1.1) n'est pas explicite, les trajectoires déterministes peuvent être simulées numériquement en utilisant un schéma d'intégration (e.g. Euler implicite, Euler explicite, Runge-Kutta).

1.1.1.1 Modèles compartimentaux classiques

Les modèles les plus basiques supposent que le mélange entre les individus est homogène et que chaque individu a la même probabilité d'être infecté et de transmettre la maladie. La représentation simple SIR (cf Figure 1.2) est une brique de base de la modélisation des dynamiques épidémiques. Tout le long de cette section, nous utilisons cette représentation comme fil directeur.

Exemple 1 : modèle SIR sans démographie Dans le modèle SIR, présenté et formalisé par Kermack et al. (1927), une première hypothèse simplificatrice est de supposer que l'épidémie évolue dans une population fermée de taille fixe N, n'incluant aucun processus démographique (e.g. naissances, décès), et où les contacts entre individus sont considérés homogènes. Le système comprend trois compartiments correspondant respectivement aux nombres d'individus susceptibles d'être infectés (S), infectieux symptomatiques (I), et retirés de la chaîne de transmission (R), assimilés ici aux immunisés après guérison. Les paramètres épidémiques du modèle sont λ le taux de transmission et γ le taux de guérison. Les dynamiques des compartiments sont décrites par le système général d'EDO (1.1) avec $d = 3$, $x = (s, i, r)^t$, $x_0 = (s_0, i_0, r_0)^t \neq (0, 0, 0)^t$ et

$$b : (s, i, r)^t \to \begin{pmatrix} -\lambda si \\ \lambda si - \gamma i \\ \gamma i \end{pmatrix}.$$

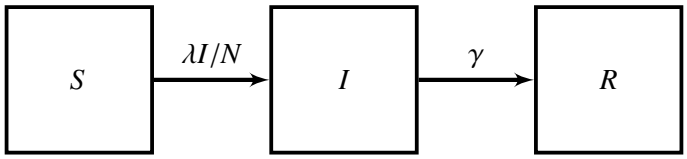

FIGURE 1.2 – Modèle à compartiments SIR avec trois blocs correspondant respectivement à des effectifs d'individus susceptibles (S), infectieux symptomatiques (I) et guéris (R). Les transitions des individus d'un état de santé à un autre sont gouvernées par le taux de transmission λ et le taux de guérison γ.

Exemple 2 : modèle SIR avec démographie Il est possible d'étendre ce premier modèle en incluant des processus démographiques tout en conservant une taille de population fixe N. Une façon simple est d'introduire un paramètre μ pour modéliser un flux entrant de naissances dans le compartiment S et un flux sortant de mortalité naturelle (*i.e.* décès survenant indépendamment de la maladie et de la pathogénécité de l'agent infectieux) dans tous les compartiments. Le modèle SIR avec démographie se formalise par le système d'EDO (1.1) avec $d = 3$, $x = (s, i, r)^t$, $x_0 = (s_0, i_0, r_0)^t \neq (0, 0, 0)^t$ et

$$b : (s, i, r)^t \to \begin{pmatrix} \mu - \lambda si - \mu s \\ \lambda si - \gamma i - \mu i \\ \gamma i - \mu r \end{pmatrix}.$$

Remarque : Le modèle SIR suppose que l'immunité d'un individu est définitive, mais il existe des situations où l'individu reste infectieux toute sa vie. Pour cela, le modèle SI est utilisé comme représentation des dynamiques d'infections durables (e.g. VIH : Kuang et al. (2007), Ghosh et al. (2018)). Une autre possibilité est de considérer que l'individu guérit mais qu'il n'y a pas d'acquisition de réponse immunitaire. Le modèle SIS permet de prendre en compte cette caractéristique

et a été utilisé pour décrire, par exemple, des infections à rotavirus (Keeling and Rohani (2007)).

A partir de ces modèles de base, des modèles plus complexes peuvent être construits en ajoutant des compartiments supplémentaires (e.g. modèle SEIR, structure en classes d'âge ou de risque ; cf Section 1.1.1.2), en considérant de nouvelles transitions entre compartiments et donc plus de paramètres et/ou en définissant des paramètres variant dans le temps (cf Section 1.1.1.3).

Exemple 3 : modèle SEIR Le modèle SEIR (cf Figure 1.3) est par exemple utilisé pour modéliser les dynamiques de la grippe saisonnière (e.g. Chowell et al. (2008), Baguelin et al. (2010), Cori et al. (2012)). Il contient un compartiment (E) qui décrit le nombre d'individus exposés (individus infectés mais pas encore infectieux). Dans ce cas, un paramètre ϵ, défini comme étant le taux d'exposition (inverse de la période d'exposition), est introduit. En notant $e = \frac{E}{N}$, les EDO du modèle SEIR (sans démographie) sont décrites dans le système (1.1) avec $d = 4$, $x = (s, e, i, r)^t$, $(s_0, e_0, i_0, r_0)^t \neq (0, 0, 0, 0)^t$ et

$$b : (s, e, i, r)^t \rightarrow \begin{pmatrix} -\lambda si \\ \lambda si - \epsilon e \\ \epsilon e - \gamma i \\ \gamma i \end{pmatrix}.$$

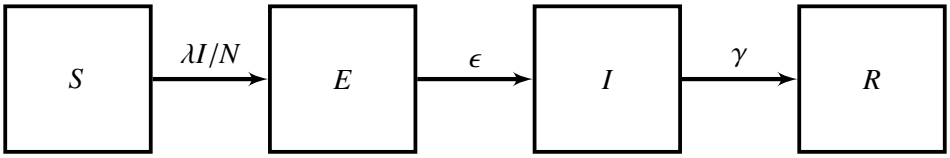

FIGURE 1.3 – Modèle à compartiments SEIR avec quatre blocs correspondant respectivement à des effectifs d'individus susceptibles (S), exposés (E), infectieux symptomatiques (I) et guéris (R). Les transitions des individus d'un état de santé à un autre sont gouvernées par le taux de transmission λ, le taux d'exposition ϵ et le taux de guérison γ.

Il est possible d'ajouter un compartiment rassemblant les individus dits porteurs sains. Dans ce cas, des individus susceptibles peuvent être infectés soit par des porteurs sains soit par des individus symptomatiques. Cette situation se produit pour des maladies telles que l'hépatite B et l'herpès (Keeling and Rohani (2007)), où certains individus infectés peuvent devenir des porteurs sains durables et transmettre l'infection à un faible taux durant plusieurs années. Pour la Covid-19, un des premiers modèles développés (Pan et al. (2020)) consiste en un modèle SEIR étendu comportant des compartiments décrivant les individus ne présentant pas de symptômes (individus *asymptomatiques*) ou avec des symptômes très légers (individus *paucisymptomatiques*) mais qui peuvent cependant transmettre la maladie.

1.1.1.2 Modèles avec structure en classes homogènes

Supposer que le mélange entre les individus est homogène dans la population est une hypothèse forte dans de nombreux contextes. Pour relâcher cette hypothèse, une solution consiste à diviser chaque compartiment en plusieurs catégories de tailles plus petites et supposer que le mélange est homogène au sein de chaque catégorie mais se fait à des taux différents entre les catégories (e.g. taux de transmission spécifiques pour modéliser les interactions entre les classes). Ainsi, en plus de la première subdivision partitionnant la population en fonction de l'état de santé des individus, se rajoute une deuxième subdivision au sein de chaque compartiment (cf Figure 1.4). Deux exemples classiques sont les structures en *classes de risque* et en *classes d'âge*.

Incorporer une structure en classes de risque est particulièrement pertinent dans le cas des maladies sexuellement transmissibles où certains individus sont dits à haut-risque lorsqu'ils ont une probabilité plus grande que les autres de transmettre la maladie (e.g. des individus ayant plusieurs partenaires sexuels dans le cas du VIH, Keeling and Rohani (2007)). Ces individus possèdent un risque plus élevé de contracter l'infection et de la transmettre. Intégrer une hétérogénéité entre individus est utile pour mieux comprendre et prédire les mécanismes de transmission de ces maladies (Johnson et al. (1994)). Des dynamiques de transmission de plusieurs maladies sexuellement transmissibles telles que la syphilis (e.g. Cates et al. (1996)), la gonorrhée (e.g. Kretzschmar et al. (1996)) et le VIH (e.g. Anderson et al. (1992)) ont été analysées à l'aide de modèles structurés en classes de risque.

La structure en classes d'âge permet de décrire des maladies fréquentes chez les enfants mais plus rares chez les adultes en introduisant des taux de contact dépendant de l'âge. Les dynamiques des infections infantiles ont beaucoup été étudiées et modélisées dans la littérature (e.g. rougeole : Ferguson et al. (1996), Bjørnstad et al. (2002), Rohani et al. (2002) ; variole : Ferguson et al. (2003) ; coqueluche : Rohani et al. (2002)).

Exemple 4 : modèle SIR avec 2 classes Considérons un modèle SIR structuré en 2 classes homogènes. Soient S_1 et S_2 (resp. I_1, I_2 et R_1, R_2) les effectifs d'individus susceptibles des groupes 1 et 2 (resp. les effectifs d'individus infectés et guéris du groupe 1 et 2). Notons $s_1 = \frac{S_1}{N}$ (resp. $i_1 = \frac{I_1}{N}$, $r_1 = \frac{R_1}{N}$, $s_2 = \frac{S_2}{N}$, $i_2 = \frac{I_2}{N}$ et $r_2 = \frac{R_2}{N}$). De plus, supposons que la transmission de la maladie entre les deux classes est décrite par la matrice des taux de transmission

$$\lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix},$$

et que les transitions entre les compartiments S_1 et S_2 (resp. entre I_1 et I_2 et entre R_1 et R_2) se font au taux l_S (resp. l_I et l_R). Par exemple, dans le cas d'une structure en classe d'âge, cela permet de tenir compte du vieillissement de la population. Cela n'a de sens que si l'intervalle de temps étudié est suffisamment long par rapport au vieillissement. De plus, dans ce cas, $l_S = l_I = l_R = l$ car le vieillissement d'un individu ne dépend pas *a priori* de son état de santé.

En considérant un taux de guérison γ indépendant du groupe, le système d'EDO décrivant les mécanismes de transmission est donné en (1.1) avec $d = 6$, $x = (s_1, i_1, r_1, s_2, i_2, r_2)^t$, $x_0 = (s_{1,0}, i_{1,0}, r_{1,0}, s_{2,0}, i_{2,0}, r_{2,0})^t \neq (0,0,0,0,0,0)^t$ et

$$b : (s_1, i_1, r_1, s_2, i_2, r_2)^t \to \begin{pmatrix} -\lambda_{11} s_1 i_1 - \lambda_{12} s_1 i_2 - l_S s_1 \\ \lambda_{11} s_1 i_1 + \lambda_{12} s_1 i_2 - \gamma i_1 - l_I i_1 \\ \gamma i_1 - l_R r_1 \\ -\lambda_{21} s_2 i_1 - \lambda_{22} s_2 i_2 + l_S s_1 \\ \lambda_{21} s_2 i_1 + \lambda_{22} s_2 i_2 - \gamma i_2 + l_I i_1 \\ \gamma i_2 + l_R r_1 \end{pmatrix}.$$

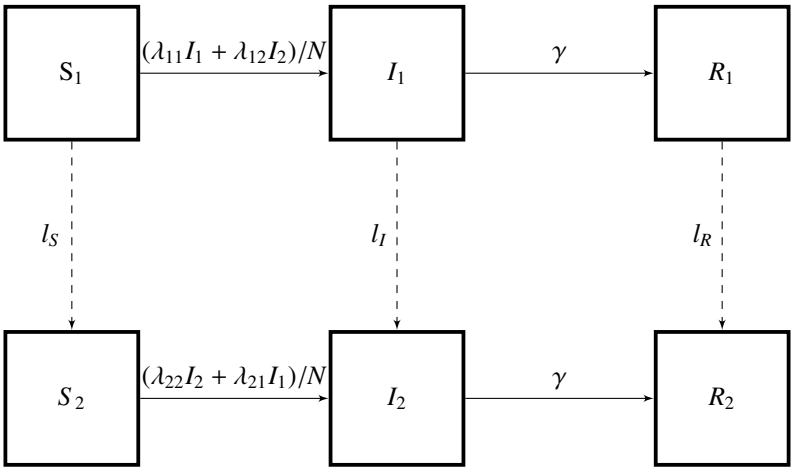

FIGURE 1.4 – Modèle à compartiments SIR structuré en 2 classes homogènes, avec l_S le taux auquel des individus susceptibles passent du groupe 1 au groupe 2 (resp. l_I et l_R le taux auquel des individus infectés et guéris passent du groupe 1 au groupe 2).

1.1.1.3 Modèles à temporalité forcée

Dans certaines situations, afin de prendre en compte des facteurs environnementaux (e.g. climat) ou sociétaux (e.g. vacances) dans les mécanismes de transmission, certains paramètres peuvent dépendre du temps. Par exemple, Finkenstädt and Grenfell (2000) ont utilisé des données de reports de cas de rougeole en Angleterre et au Pays de Galles de 1944 à 1964 pour estimer des taux de transmission bihebdomadaires.

Exemple 5 : modèle SIR avec un taux de transmission non-homogène dans le temps Pour modéliser une composante temporelle dans le processus de transmission de la rougeole, Bailey (1975) est un des premiers à incorporer dans les équations du modèle SIR un taux de transmission non-homogène dans le temps $\lambda(t)$ exprimé par une fonction sinusoïdale :

$$\lambda(t) = \lambda_0(1 + \lambda_1 \cos(\omega t)),$$

où λ_0 correspond à un taux de transmission moyen, ω est la période du forçage sinusoïdal et λ_1 décrit l'amplitude de la saisonnalité dans la transmission. Cela implique non pas une modification de la structure compartimentale du modèle, mais une modification de la paramétrisation de la fonction $b(\cdot)$ dans (1.1) :

$$b : (t, (s, i, r)^t) \to \begin{pmatrix} -\lambda(t)si \\ \lambda(t)si - \gamma i \\ \gamma i \end{pmatrix}.$$

1.1.1.4 Modélisation de mesures de contrôle

Il est possible d'incorporer dans les modèles épidémiologiques des statégies de contrôle afin d'en étudier l'impact sur la transmission au sein d'une population. Un premier exemple de stratégie est la vaccination. Celle-ci peut agir en réduisant la probabilité d'être infecté, la transmissibilité dans le cas d'une infection ou la durée de la phase infectieuse. Un moyen de prendre en compte la vaccination est d'introduire un paramètre ν exprimé comme le produit entre le taux de vaccination et la probabilité de développer une immunité après injection du vaccin et/ou d'ajouter un compartiment supplémentaire correspondant aux individus vaccinés.

Exemple 6 : modèle SIR avec vaccination Le système d'EDO (1.1) formalise les dynamiques des effectifs des compartiments avec prise en compte de la vaccination (supposée parfaite) à travers un paramètre ν, où $d = 3$, $x = (s, i, r)^t$, $x_0 = (s_0, i_0, r_0)^t \neq (0, 0, 0)^t$ et

$$b : (s, i, r)^t \to \begin{pmatrix} -\lambda si - \nu s \\ \lambda si - \gamma i \\ \gamma i + \nu s \end{pmatrix}.$$

D'une manière alternative, dans l'objectif d'analyser des dynamiques de la Covid-19 en France tout en prenant en compte la couverture vaccinale, Collin et al. (2021) ont ajouté un compartiment V pour décrire le nombre d'individus vaccinés en fonction du temps.

Une autre mesure efficace concerne l'isolement d'individus, celle-ci pouvant se concentrer sur certains types d'individus seulement, mais aussi sur toute la population. Une façon simple de modéliser une mesure de quarantaine consiste à ajouter un compartiment supplémentaire Q réunissant les individus infectieux détectés puis isolés.

Exemple 7 : modèle SIR avec isolement Soit ζ le taux auquel les individus infectieux sont détectés puis isolés et $1/\omega$ la durée moyenne de l'isolement. En notant $q = \frac{Q}{N}$, le système d'EDO incorporant une mesure d'isolement s'exprime sous la forme ((1.1), cf Figure 1.5), avec $d = 4$, $x = (s, i, r, q)^t$, $(s_0, i_0, r_0, q_0)^t \neq (0, 0, 0, 0)^t$ et

$$b : (s, i, r, q)^t \to \begin{pmatrix} -\lambda si \\ \lambda si - \gamma i - \zeta i \\ \gamma i + \omega q \\ \zeta i - \omega q \end{pmatrix}.$$

A titre d'exemple, des mesures d'isolement d'individus infectés et des cas contact ont été instaurés par le gouvernement français pour réduire la transmission de la Covid-19. De plus, de nombreuses études considèrent un compartiment décrivant le nombre de personnes hospitalisées en fonction du temps (voir e.g. Prague et al. (2020), Collin et al. (2021), Cazelles et al. (2021)) et qui peut être assimilé à une mesure d'isolement sous l'hypothèse que les individus hospitalisés ne peuvent plus transmettre la maladie.

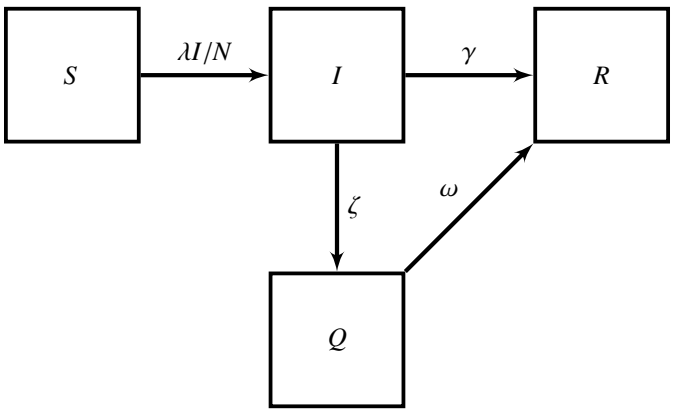

FIGURE 1.5 – Modèle à compartiments SIR avec mesure d'isolement des individus infectieux dans le compartiment Q.

1.1.1.5 Modélisation avec perte d'immunité

Certaines maladies induisent une immunité après infection qui est seulement temporaire ou imparfaite, une raison pouvant être l'évolution de l'agent pathogène. Ces phénomènes peuvent

conduire à des dynamiques récurrentes, comme dans le cas de la grippe saisonnière (Cauchemez et al. (2008)).

Exemple 8 : modèle SIRS Le modèle SIRS est utilisé pour décrire des infections pour lesquelles l'immunité après guérison est temporaire. Il peut s'exprimer sous la forme ((1.1), cf Figure 1.6) avec $d = 3$, $x = (s, i, r)^t$, $(s_0, i_0, r_0)^t \neq (0, 0, 0)^t$ et :

$$b : (s, i, r)^t \to \begin{pmatrix} -\lambda si + ur \\ \lambda si - \gamma i \\ \gamma i - ur \end{pmatrix},$$

où u est le taux auquel l'immunité est perdue, cela entraînant un retour des individus du compartiment R au compartiment S.

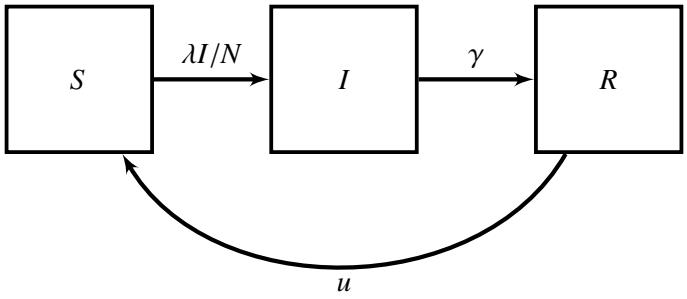

FIGURE 1.6 – Modèle à compartiments SIRS, avec un paramètre u décrivant le taux auquel l'immunité après infection est perdue.

1.1.1.6 Modélisation d'épidémies multisites

Dans certaines circonstances, il est naturel de considérer conjointement les mécanismes de transmission d'une épidémie multisite regroupant des populations d'individus séparées spatialement (e.g. SARS-CoV-2 dans plusieurs régions de France). En particulier, les modèles de métapopulation sont des modèles mécanistes incorporant une structure de connectivité dans le processus de transmission (Keeling and Rohani (2007)). Dans ces modèles, la population totale est divisée en plusieurs *sous-populations* possédant chacune des dynamiques indépendantes et avec une interaction limitée entre les sous-populations. Par exemple, en cas de proximité géographique, la transmission d'une maladie aéroportée peut se faire à travers des petites particules suspendues dans l'air. D'une façon plus générale, la transmission entre les sous-populations se produit principalement par des mouvements d'individus infectés entre les sous-populations.

Exemple 9 : modèle SIR de métapopulation Soient J le nombre de sous-populations et s_j, $1 \leq j \leq J$, (resp. i_j, r_j) la proportion d'individus susceptibles (resp. la proportion d'individus infectés et guéris) se trouvant dans la sous-population j. Nous notons m_{jl} le taux auquel les individus se déplacent d'une sous-population j à une sous-population l. Un modèle SIR avec démographie de métapopulation, prenant en compte des déplacements des individus, peut s'exprimer sous la forme (1.1) avec $d = 3 \times J$ et, pour tout $j \in \{1, \ldots, J\}$, $x_j = (s_j, i_j, r_j)^t$, $(s_{0,j}, i_{0,j}, r_{0,j})^t \neq (0, 0, 0)^t$ et

$$b : (s_j, i_j, r_j)^t \to \begin{pmatrix} \mu_j - \lambda_j s_j i_j - \mu_j s_j + \sum_{l=1}^{J} m_{lj} s_l - \sum_{l=1}^{J} m_{jl} s_j \\ \lambda_j s_j i_j - \gamma_j i_j - \mu_j i_j + \sum_{l=1}^{J} m_{lj} i_l - \sum_{l=1}^{J} m_{jl} i_j \\ \gamma_j i_j - \mu_j r_j + \sum_{l=1}^{J} m_{lj} r_l - \sum_{l=1}^{J} m_{jl} r_j \end{pmatrix}.$$

Les paramètres ont la même signification que dans l'Exemple 2, Section (1.1.1.1), mais sont définis à l'échelle de la sous-population.

1.1.2 Modèles stochastiques de propagation d'épidémies

La propagation des maladies infectieuses est influencée par plusieurs sources de stochasticité. Une première source est la stochasticité démographique (ou intrinsèque) qui rend compte du fait que tous les individus d'une population ne réagissent pas de la même manière face à la maladie. L'effet stochastique de la démographie sur l'évolution de l'épidémie est d'autant plus important que le nombre d'individus infectieux est faible, par exemple lorsque la taille de population est petite (Keeling and Rohani (2007)) ou en début et fin d'épidémie. Une deuxième source est la stochasticité environnementale, regroupant toutes les perturbations externes à la population ayant un impact sur les mécanismes épidémiques (e.g. température, humidité).

Les modèles stochastiques sont considérés comme des représentations plus réalistes et naturelles que leur équivalent déterministe pour prendre en compte les différentes sources de stochasticité, bien que leur analyse mathématique soit souvent aussi plus délicate (Andersson and Britton (2000), Breto et al. (2009), Britton and Pardoux (2020)). Par exemple, l'utilisation de processus stochastiques décrivant des dynamiques démographiques et/ou environnementales a permis un meilleur ajustement des modèles par rapport aux dynamiques observées (e.g. rougeole : He et al. (2010)). Par une étude sur simulations suivie d'une analyse sur une phase précoce de l'épidémie du virus Ebola en 2014 en Afrique de l'Ouest, King et al. (2015) ont montré que l'utilisation de modèles stochastiques plutôt que déterministes dans une procédure d'inférence entraîne une réduction du biais des paramètres estimés ainsi qu'une meilleure quantification de l'incertitude dans les estimations et les prédictions.

Une première approche stochastique consiste à garder la structure en compartiments contenant les différents états de santé et de supposer un processus de comptage multidimensionnel décrivant à chaque instant t le nombre d'individus dans chaque compartiment. De nombreux auteurs, tenant compte du fait que les observations sont recueillies à des temps discrets, ont utilisé le cadre des séries chronologiques (Breto et al. (2009)). Ce cadre stochastique étant très large, des modèles stochastiques obtenus avec des hypothèses plus fortes ont été proposés.

Pour le début d'une épidémie, une approximation courante en grande population consiste à supposer une taille de population infinie d'individus susceptibles, la courbe d'infectés suivant alors un processus de branchement (Britton (2010)). Cela permet en particulier d'obtenir facilement une expression pour la probabilité d'observer une épidémie majeure (*i.e.* ne s'éteignant pas précocement).

Dans cette thèse, nous nous intéressons à la modélisation stochastique d'épidémies majeures dans des populations fermées de taille N sur un intervalle de temps $[0, T]$ via les processus Markoviens de sauts et leurs approximations (e.g. Britton and Pardoux (2020), Ethier and Kurtz (2005)).

1.1.2.1 Processus Markoviens de sauts

Considérons une épidémie décrite par d stades d'infection (ou compartiments), les effectifs dans chaque compartiment étant régis par un processus Markovien de sauts

$$\mathcal{Z}(t) = (\mathcal{Z}_1(t), \dots, \mathcal{Z}_d(t))^t,\ t \geq 0, \tag{1.2}$$

à espace d'état $E = \{0, \dots, N\}^d$ et de matrice de transition $Q = (q_{j,k}),\ j, k \in E$, telle que pour tout $j \neq k$,

$$q_{j,k} \geq 0 \text{ et } q_{j,j} = -\sum_{j \in E, k \neq j} q_{j,k}.$$

Soit $\alpha_l(.) : E \to \mathbb{R}^+$ la collection de fonctions décrivant les sauts d'amplitude l, avec $l \neq (0,\dots,0)^t$:

$$\alpha_l(k) = q_{k,k+l}.$$

Ces fonctions vérifient :

$$\forall k \in E,\ 0 < \sum_{l \in E} \alpha_l(k) = \alpha(k) < +\infty.$$

Ainsi, le processus stochastique $\mathcal{Z}(t)$ demeure dans chaque état $k \in E$ durant un temps exponentiel $\mathcal{E}(\alpha(k))$ et saute à l'état $k + l$ avec probabilité $\frac{\alpha_l(k)}{\alpha(k)}$.

Exemple 10 : modèle SIR Soit $\mathcal{Z}(t) = (S(t), I(t))^t$ un processus Markovien de sauts à espace d'état $E = \{0,\dots,N\}^2$ (nous omettons le compartiment des guéris car $R(t) = N - S(t) - I(t)$ pour tout t). Seuls deux sauts sont possibles à partir de $k = (S, I)^t$:

— $\ell_1 = (-1, +1)^t : (S, I) \to (S - 1, I + 1) \Rightarrow q_{k,k+\ell_1} = \lambda S I/N = \alpha_{\ell_1}(k)$ (infection),

— $\ell_2 = (0, -1)^t : (S, I) \to (S, I - 1) \Rightarrow q_{k,k+\ell_2} = \gamma I = \alpha_{\ell_2}(k)$ (guérison).

Exemple 11 : modèle SIR avec isolement Nous reprenons l'exemple 7. Soit $\mathcal{Z}(t) = (S(t), I(t), Q(t))^t$ un processus Markovien de sauts à espace d'état $E = \{0,\dots,N\}^3$ avec $Q(t)$ le nombre d'individus isolés et ne pouvant plus transmettre la maladie. Quatre sauts sont possibles à partir de $k = (S, I, Q)^t$:

— $\ell_1 = (-1, +1, 0)^t : (S, I, Q) \to (S - 1, I + 1, Q) \Rightarrow q_{k,k+\ell_1} = \lambda S I/N = \alpha_{\ell_1}(k)$ (infection),

— $\ell_2 = (0, -1, 0)^t : (S, I, Q) \to (S, I - 1, Q) \Rightarrow q_{k,k+\ell_2} = \gamma I = \alpha_{\ell_2}(k)$ (guérison des individus infectés mais pas isolés),

— $\ell_3 = (0, -1, +1)^t : (S, I, Q) \to (S, I - 1, Q + 1) \Rightarrow q_{k,k+\ell_3} = \zeta I = \alpha_{\ell_3}(k)$ (isolement),

— $\ell_4 = (0, 0, -1)^t : (S, I, Q) \to (S, I, Q - 1) \Rightarrow q_{k,k+\ell_4} = \omega Q = \alpha_{\ell_4}(k)$ (guérison des individus infectés et isolés).

Remarques :

- La propriété d'absence de mémoire des processus Markoviens de sauts implique que les durées de séjour dans les compartiments suivent une distribution exponentielle. Cette hypothèse simplificatrice facilite grandement l'analyse mathématique de ces modèles, bien qu'elle n'ait pas réellement de justification épidémiologique (Andersson and Britton (2000)) et qu'elle soit peu plausible pour un certain nombre de maladies infectieuses (e.g. syndrome respiratoire aigu sévère (SRAS) : Donnelly et al. (2003) ; encéphalopathie spongiforme bovine (ESB) : Ferguson et al. (1997)). Certains travaux spécifient d'autres distributions pour les durées de séjour dans les compartiments (e.g. distribution Gamma : Donnelly et al. (2003), Ferguson et al. (1997), distribution de Weibull : Ferguson et al. (1997)). Une alternative consiste à étendre l'espace d'état en définissant des sous-compartiments (e.g. compartiment I divisé en m compartiments $I_1,\dots,I_m$) avec des durées de séjour distribuées exponentiellement. Dans ce cas, la somme de ces durées de séjour correspondant à la durée globale passée dans l'état infectieux suit une loi d'Erlang (somme de variables aléatoires exponentielles indépendantes et identiquement distribuées (i.i.d), Lloyd (2001)).
- Le processus de sauts défini en (1.2) peut être simulé de façon exacte par l'algorithme de Gillespie (Gillespie (1977)). A partir d'un état donné du système, on simule le temps jusqu'au prochain saut selon une loi exponentielle de paramètre la somme des taux de transition entre les compartiments, puis on sélectionne uniformément quel saut (*i.e.* transition) s'est produit. Comme tous les sauts sont générés, le coût de calcul peut être élevé lorsque la taille

de population est grande. Gillespie (2001) a proposé une approximation, la méthode τ-leap, où les sauts se font à des temps discrets espacés par un pas de temps τ et les taux de transition densité-dépendants sont constants entre deux instants de saut t et $t + \tau$. La qualité de cette approximation est d'autant meilleure que la valeur choisie pour τ est faible, mais un compromis doit être trouvé pour obtenir des temps de calcul raisonnables.

Il existe des algorithmes basés sur des formalisations différentes du processus de sauts via la limite de processus multinomiaux couplés et discrets en temps avec des taux aléatoires. Dans ce cas, on peut utiliser des schémas numériques d'intégration tels que le schéma d'Euler multinomial (Breto et al. (2009)) dans lequel les nombres de transfert d'individus entre deux compartiments durant un intervalle de temps $[t, t + \delta]$, avec δ un pas de temps discret donné, suivent des distributions multinomiales.

1.1.2.2 Processus Markoviens de sauts densité-dépendants

Notons $(\mathcal{Z}_N(t))_{t\geq 0}$ le processus Markovien de sauts normalisé par la taille de population N :

$$\mathcal{Z}_N(t) = \frac{\mathcal{Z}(t)}{N} \in E^N = \{l/N, l \in E\}. \tag{1.3}$$

Pour $x \in E^N$, les fonctions de saut associées à ce processus sont notées $\alpha_l^N(x) = \frac{1}{N}\alpha_l([Nx])$. Le processus $\mathcal{Z}(t)$ défini en (1.2) est densité-dépendant s'il vérifie les hypothèses suivantes :

$$\textbf{H1 :} \forall l \in E,\ \forall x \in [0,1]^d,\ \ \alpha_l^N(x) \underset{N\to+\infty}{\to} \beta_l(x),$$
$$\textbf{H2 :} \forall l \in E,\ \beta_l \in C^2([0,1]^d),$$

où $\beta_l : [0,1]^d \to \mathbb{R}^+$ est une collection de fonctions et $[Nx] = ([Nx_1], \ldots, [Nx_d])$ avec $[Nx_i]$ correspondant à la partie entière de Nx_i.

Exemple 12 : modèle SIR Soient $s = \frac{S}{N}$ et $i = \frac{I}{N}$. Alors, quand $N \to \infty$:

$$\frac{1}{N}\alpha_{\ell_1}([Ns],[Ni]) = \frac{1}{N}(\lambda[Ns])\frac{[Ni]}{N} \to \lambda si,$$
$$\frac{1}{N}\alpha_{\ell_2}([Ns],[Ni]) = \frac{1}{N}\gamma[Ni] \to \gamma i.$$

Exemple 13 : modèle SIR avec isolement Soient $s = \frac{S}{N}$, $i = \frac{I}{N}$ et $q = \frac{Q}{N}$. Alors, quand $N \to \infty$:

$$\frac{1}{N}\alpha_{\ell_1}([Ns],[Ni],[Nq]) = \frac{1}{N}(\lambda[Ns])\frac{[Ni]}{N} \to \lambda si,$$
$$\frac{1}{N}\alpha_{\ell_2}([Ns],[Ni],[Nq]) = \frac{1}{N}\gamma[Ni] \to \gamma i,$$
$$\frac{1}{N}\alpha_{\ell_3}([Ns],[Ni],[Nq]) = \frac{1}{N}\zeta[Ni] \to \zeta i,$$
$$\frac{1}{N}\alpha_{\ell_4}([Ns],[Ni],[Nq]) = \frac{1}{N}\omega[Nq] \to \omega q.$$

Remarque : Le lien entre la modélisation déterministe par des EDO et les processus Markoviens de sauts densité-dépendants s'obtient à travers la loi des grands nombres (voir e.g. Britton and Pardoux (2020), Chapitre 2). Supposons que $(\mathcal{Z}(t))$ satisfait les hypothèses (**H1**), (**H2**), et $\mathcal{Z}_N(0) \to x_0$ quand $N \to +\infty$. Alors, $(\mathcal{Z}_N(t))$ converge presque sûrement uniformément sur $[0,T]$ vers la solution $x(t)$ de l'équation différentielle ordinaire (1.1).

A partir des processus Markoviens de sauts densité-dépendants, plusieurs approximations peuvent être obtenues : les processus de diffusion et les processus Gaussiens.

1.1.2.3 Processus de diffusion en grande population

Les processus de diffusion ont été considérés comme approximations des dynamiques épidémiques (Ross et al. (2009), Fuchs (2013), Guy et al. (2014), Guy et al. (2015)) et utilisés dans plusieurs études épidémiologiques (e.g. choléra : King et al. (2008) ; SARS-CoV-2 : Cazelles et al. (2021)). A la différence des processus Markoviens de sauts, les processus de diffusion sont à espace d'état continu. L'approximation, reposant sur plusieurs hypothèses simplificatrices, nécessite que $\mathcal{Z}_N(0) \to x_0 \neq (0,\dots,0)^t$ quand $N \to +\infty$ (*i.e.* grande population).

Soient $b : [0,1]^d \to \mathbb{R}^d$ la fonction de dérive et Σ une matrice symétrique positive de taille $d \times d$ vérifiant pour tout x dans $[0,1]^d$:

$$b(x) = \sum_{\ell \in E^-} \ell\, \beta_\ell(x)\ ; \quad \Sigma(x) = \sum_{\ell \in E^-} \beta_\ell(x)\, \ell\, \ell^t. \tag{1.4}$$

Il est à noter que $b(\cdot)$ et $\Sigma(\cdot)$ peuvent dépendre de paramètres et du temps que nous omettons par souci de clarté. En se basant sur les résultats de Ethier and Kurtz (2005), Guy et al. (2015) ont proposé d'approximer les processus de sauts par des processus de diffusion multidimensionnels $(Z_N(t))_{t\geq 0}$ à petite variance $\frac{1}{N}\Sigma(x)$, où Σ est la matrice définie dans (1.4) :

$$\begin{cases} dZ_N(t) &= b(Z_N(t)) + \frac{1}{\sqrt{N}}\sigma(Z_N(t))\, dB(t), \\ Z_N(0) &= x_0, \end{cases} \tag{1.5}$$

avec $(B(t))_{t\geq 0}$ un mouvement Brownien d-dimensionnel et σ une matrice de taille $d \times d$ telle que

$$\sigma(x)\sigma^t(x) = \Sigma(x). \tag{1.6}$$

Exemple 14 : modèle SIR Les quantités définies en (1.4) et (1.6) sont respectivement :

$$b(s,i) = \begin{pmatrix} -\lambda si \\ \lambda si - \gamma i \end{pmatrix}, \quad \Sigma(s,i) = \begin{pmatrix} \lambda si & -\lambda si \\ -\lambda si & \lambda si + \gamma i \end{pmatrix}, \quad \sigma(s,i) = \begin{pmatrix} \sqrt{\lambda si} & 0 \\ -\sqrt{\lambda si} & \sqrt{\gamma i} \end{pmatrix}. \tag{1.7}$$

Soit $Z_N(t) = (S_N(t), I_N(t))^t$ le processus de diffusion à petite variance. En utilisant (1.7), les équations différentielles stochastiques vérifient :

$$\begin{cases} dS_N(t) &= -\lambda S_N(t) I_N(t) + \frac{1}{\sqrt{N}}\sqrt{\lambda S_N(t) I_N(t)}\, dB_1(t), \\ dI_N(t) &= (\lambda S_N(t) I_N(t) - \gamma I_N(t)) - \frac{1}{\sqrt{N}}\left(\sqrt{\lambda S_N(t) I_N(t)}\, dB_1(t) - \sqrt{\gamma I_N(t)}\, dB_2(t)\right), \end{cases}$$

avec $B(t) = (B_1(t), B_2(t))^t$ un mouvement Brownien bidimensionnel et $(S_N(0), I_N(0))^t = (s_0, i_0)^t \neq (0,0)^t$.

Exemple 15 : modèle SIR avec isolement Les quantités définies en (1.4) et (1.6) sont respectivement :

$$b(s,i,q) = \begin{pmatrix} -\lambda si \\ \lambda si - \gamma i - \zeta i \\ \zeta i - \omega q \end{pmatrix}, \quad \Sigma(s,i,q) = \begin{pmatrix} \lambda si & -\lambda si & 0 \\ -\lambda si & \lambda si + (\gamma+\zeta) i & -\zeta i \\ 0 & -\zeta i & \zeta i + \omega q \end{pmatrix},$$

et

$$\sigma(s,i,q) = \begin{pmatrix} \sqrt{\lambda si} & 0 & 0 \\ -\sqrt{\lambda si} & \sqrt{(\gamma+\zeta) i} & 0 \\ 0 & -\frac{\zeta i}{\sqrt{(\gamma+\zeta) i}} & \sqrt{\zeta i + \omega q - \frac{\zeta^2 i^2}{(\gamma+\zeta) i}} \end{pmatrix}.$$

1.1.2.4 Approximation Gaussienne

Pour des équations différentielles stochastiques à petite variance, une approximation du processus de diffusion $Z_N(t)$ peut être obtenue en utilisant (1.4)-(1.6), basée sur la théorie des petites perturbations des systèmes dynamiques (e.g. Azencott (1982), Freidlin and Wentzell (1978)) :

$$\begin{cases} Z_N(t) & = x(t) + \frac{1}{\sqrt{N}} g(t) + \frac{1}{\sqrt{N}} R_N(t), \\ dg(t) & = \nabla_x b(x(t))\ g(t)\ dt + \sigma(x(t))\ dB(t)\,; \quad g(0) = 0, \\ \text{avec} & \sup_t \|R_N(t)\| \to 0 \ \text{ en probabilité quand } N \to +\infty, \end{cases} \tag{1.8}$$

où $\nabla_x b(x)$ est la matrice $(\frac{\partial b_i}{\partial x_j}(x))_{1\le i,j\le d}$. L'équation différentielle stochastique pour $g(\cdot)$ définie en (1.8) peut être résolue explicitement (voir e.g. Guy et al. (2014) pour plus de détails) et sa solution est le processus Gaussien inhomogène en temps vérifiant

$$g(t) = \int_0^t \Phi(t, s)\sigma(x(s))\ dB(s), \tag{1.9}$$

où $\Phi(t, s)$ vérifie l'EDO

$$\frac{\partial \Phi}{\partial t}(t, s) = \nabla_x b(x(t))\Phi(t, s), \Phi(s, s) = I_d,$$

avec I_d désignant la matrice identité. Ainsi, $\Phi(t, s)$ est la matrice de taille $d \times d$

$$\Phi(t, s) = \exp\left(\int_s^t \nabla_x b(x(u))\ du\right).$$

En utilisant (1.1) et (1.9), le processus Gaussien $G_N(t)$ est défini comme suit (Guy et al. (2015)) :

$$G_N(t) = x(t) + \frac{1}{\sqrt{N}} g(t). \tag{1.10}$$

1.1.2.5 Conséquences statistiques

Pour les études statistiques, il est possible de travailler sur l'un des trois processus présentés : le processus Markovien de sauts $\mathcal{Z}_N(t)$ (1.3), le processus de diffusion $Z_N(t)$ (1.5) ou le processus Gaussien $G_N(t)$ (1.10).

En effet, considérons la distance Wasserstein-1 sur l'intervalle de temps $[0, T]$ entre deux processus à valeurs dans $\mathbb{R}^d$ et définis sur $[0, T]$, notés U_t et V_t. Cette distance vérifie :

$$W_{1,T}(U, V) = \inf \mathbb{E}(\|U - V\|_T),$$

avec, si $x : [0, T] \to \mathbb{R}^d$, $\|x\|_T = \sup_{0\le t\le T} \|x(t)\|$. Dans Britton and Pardoux (2020), Chapitre 2, les auteurs ont montré que pour tout $T > 0$, les distances Wasserstein-1 sur $[0, T]$ entre les trois processus $(\mathcal{Z}_N(\cdot))$, $(Z_N(\cdot))$, et $(G_N(\cdot))$ définies en (1.3),(1.5),(1.10) satisfont, quand $N \to \infty$:

$$\sqrt{N} W_{1,T}(\mathcal{Z}_N, Z_N) \to 0, \quad \sqrt{N} W_{1,T}(\mathcal{Z}_N, G_N) \to 0, \quad \text{et } \sqrt{N} W_{1,T}(Z_N, G_N) \to 0.$$

Dans cette thèse, nous utiliserons le processus Gaussien $G_N(t)$ pour approximer les dynamiques épidémiques qui possède des propriétés intéressantes pour le développement de méthodes d'inférence. En particulier, dans un contexte où les données épidémiques sont recueillies à des temps discrets, la discrétisation de ce processus Gaussien nous permettra de nous placer dans le cadre des modèles à espace d'état Gaussiens.

Remarques :

- Tous les processus épidémiques décrits précédemment dépendent de paramètres que nous avons omis par souci de clarté. Par exemple, le processus Gaussien défini en (1.10) s'écrit : $G_N(\eta, t)$, où η est un vecteur de paramètres.
- Les modèles mécanistes (cf exemples (1)-(9)) ont été présentés dans un formalisme déterministe mais il existe une représentation stochastique pour tous ces modèles (certains modèles seront détaillés dans les chapitres suivants).
- Nous avons défini la fonction de dérive et la matrice de diffusion (1.4) de façon homogène dans le temps. Il est tout à fait possible de considérer leur version inhomogène dans le temps, par exemple dans le cas où le taux de transmission dépend du temps (cf Exemple 5).

1.2 Modélisation des observations

Dans l'objectif de fixer un cadre statistique permettant de faire de l'inférence de paramètres, nous présentons différentes modélisations possibles des observations. Comme mentionné dans le préambule, les observations disponibles sur l'évolution d'épidémies sont souvent incomplètes et affectées d'erreurs et autres sources de variabilité. Ces spécificités doivent être prises en compte dans la modélisation statistique. Il est à souligner que les étapes de modélisation mécaniste et de modélisation des observations ne sont pas indépendantes ni nécessairement opérées de façon successive. En effet, les particularités des observations peuvent aussi influer sur le choix du modèle mécaniste afin de faciliter l'inférence des paramètres.

Une première spécificité concerne la partialité des données en temps et en nombre de coordonnées du système. En effet, les observations disponibles sont souvent accessibles de façon quotidienne ou hebdomadaire et ne portent généralement que sur certains compartiments du système épidémique. Dans la suite, nous notons $X(t)$ un processus stochastique générique à temps continu et de dimension d décrivant l'évolution au cours du temps des différents compartiments épidémiques. Par exemple, il peut être assimilé à un des trois processus stochastiques présentés précédemment (e.g. le processus Markovien de sauts $\mathcal{Z}_N(\cdot)$ (1.3), le processus de diffusion $Z_N(\cdot)$ (1.5) ou le processus Gaussien $G_N(\cdot)$ (1.10)). La difficulté ici est que le processus $X(t)$ est à temps continu, ce qui nécessite de définir une représentation à temps discret de l'évolution de l'épidémie.

Nous considérons une discrétisation en temps du processus épidémique $X(t)$ et nous notons $\Delta > 0$ le pas de temps séparant deux observations consécutives et $t_k = k\Delta$, $0 \leq k \leq n$, les instants d'observation avec n le nombre total d'observations. Le dernier instant $T = n\Delta$ correspond à la dernière date de report des données et recouvre la fin de l'épidémie.

Nous utilisons le formalisme des modèles à espace d'état (Cappé et al. (2005)), où les états cachés/latents sont les d composantes du processus épidémique $X(\cdot)$ et les observations une séquence de variables aléatoires dont la distribution dépend de ces états latents.

En posant $X_k := X(t_k)$ et en notant $Y_k := Y(t_k)$ les observations, la modélisation de la dynamique épidémique et de ses observations prend donc la forme (cf Figure 1.1) :

$$\begin{cases} X_k|(X_{k-1};\eta) & \sim f(X_{k-1};\eta),\ k = 1,\dots,n, & \text{(1.11a)} \\ X_0 & \sim f(X_0;\eta), & \text{(1.11b)} \\ Y_k|(X_k;\mu) & \sim g(X_k;\mu),\ k = 0,\dots,n, & \text{(1.11c)} \end{cases}$$

où $f(\cdot)$ et $g(\cdot)$ désignent des distributions de probabilité, η est l'ensemble des paramètres régissant la dynamique épidémique et μ est l'ensemble des paramètres du modèle d'observation.

La forme de l'équation (1.11a) découle de la discrétisation du processus $X(t)$ en temps continu. Les sections suivantes décrivent différentes façons d'écrire l'équation (1.11c) en fonction des caractéristiques des observations disponibles.

1.2.1 Modélisation des données de prévalence et d'incidence

Il est important de bien distinguer les dynamiques d'incidence et de prévalence. Pour rappel, les données de prévalence mesurent le nombre d'individus dans un certain état de santé par rapport à la maladie à un instant donné tandis que les données d'incidence dénombrent les nouveaux cas entre deux instants consécutifs.

1.2.1.1 Prévalence

Les coordonnées $(X_k)_{k\geq 0}$ du système ne sont pas toutes observées. Formellement, cela revient à introduire un opérateur de projection $B : \mathbb{R}^d \to \mathbb{R}^q$, $q \leq d$, tel que BX_k contient uniquement les composantes observables au temps t_k. En l'occurrence, B est une matrice de taille $d \times q$ ayant pour éléments les valeurs 0 et 1. Conditionnellement à X_k, la distribution des observations de prévalence est telle que :

$$\mathbb{E}(Y_k|X_k = x_k) = Bx_k,\ k = 0, \dots, n. \tag{1.12}$$

Exemple 16 : données de prévalence dans le modèle SIR Soit $X_k = (S_k, I_k)'$ avec $k \geq 0$. Conditionnellement à $X_k = x_k$, le nombre moyen d'individus infectés au temps t_k s'écrit sous la forme (1.12) avec $B = (0 \quad 1)$.

Remarque : Il est à noter que l'opérateur de projection B peut aussi dépendre des paramètres du modèle (cf chapitres suivants) et donc inclure des erreurs de report. Dans la suite de l'introduction, pour simplifier, nous n'en considérons pas.

1.2.1.2 Incidence

Dans beaucoup d'études d'inférence, les observations utilisées concernent principalement les nombres de nouveaux cas (voir e.g. Stocks et al. (2018), Stocks (2017), Bretó (2018)). Pour définir des données d'incidence dans un formalisme mathématique, l'idée est d'introduire une quantité $\Delta_k X := X(t_k) - X(t_{k-1})$ correspondant aux accroissements de X au temps t_k.

Ainsi, conditionnellement à $\Delta_k X = \Delta_k x$, la distribution des observations d'incidence est telle que :

$$\mathbb{E}(Y_k|\Delta_k X = \Delta_k x) = B\Delta_k x,\ k = 0, \dots, n. \tag{1.13}$$

Remarques :

- On peut réécrire les équations (1.11a) et (1.11b) sur les accroissements $(\Delta_k X)_{k\geq 1}$:

$$\begin{cases} \Delta_k X|(\Delta_1 X, \dots, \Delta_{k-1} X; \eta) & \sim \tilde{f}(\Delta_1 X, \dots, \Delta_{k-1} X; \eta),\ k = 1, \dots, n, \\ X_0 & \sim f(X_0; \eta), \end{cases} \tag{1.14}$$

  où $\tilde{f}(\cdot)$ désigne une distribution de probabilité. C'est ce que nous ferons dans le **Chapitre 3** pour faciliter la mise en oeuvre de techniques de filtrage pour l'inférence, la subtilité étant que contrairement à $(X_k)_{k\geq 0}$, le processus $(\Delta_k X)_{k\geq 1}$ n'est pas Markovien car il dépend de

toute l'information passée. En utilisant (1.14), le système (1.11) se réécrit alors (cf Figure 1.8) :

$$\begin{cases} \Delta_k X | (\Delta_1 X, \dots, \Delta_{k-1} X; \eta) & \sim \tilde{f}(\Delta_1 X, \dots, \Delta_{k-1} X; \eta), \ k = 1, \dots, n, & \text{(1.15a)} \\ X_0 & \sim f(X_0; \eta), & \text{(1.15b)} \\ Y_k | (\Delta_k X; \mu) & \sim \tilde{g}(\Delta_k X; \mu), \ k = 1, \dots, n, & \text{(1.15c)} \end{cases}$$

avec $\tilde{g}(\cdot)$ désignant une densité de probabilité.

- Si la condition initiale X_0 est connue, connaître $(\Delta_k X)_{k \geq 1}$ revient à directement connaître $(X_k)_{k \geq 0}$. Autrement dit, le passage entre la prévalence et l'incidence est explicite. En pratique, comme la valeur de X_0 est rarement disponible, ces deux types de données sont à considérer. En utilisant le modèle simple SIR, Park and Bolker (2020) ont montré que faire des hypothèses incorrectes sur le temps des reports d'incidence, *i.e.* considérer que les dynamiques d'incidence et de prévalence sont confondues, entraîne un biais dans les estimations des paramètres et des intervalles de confiance trop resserrés. Cependant, si le pas de temps du report des cas et le temps de génération de la maladie G_T (*i.e.* la durée entre le moment où un individu est infecté et celui où ce même individu transmet la maladie à quelqu'un d'autre) ont des valeurs proches, alors il est attendu que les dynamiques d'incidence et de prévalence soient elles aussi similaires (Fine and Clarkson (1982)). Pour illustrer cela, nous nous plaçons dans le cadre du modèle SIR avec $\lambda = 0.6$, $\gamma = 0.4$, $\left(\frac{S(0)}{N}, \frac{I(0)}{N}, \frac{R(0)}{N}\right) = (0.95, 0.01, 0.04)$ et pour une taille de population $N = 10,000$ (cf Figure 1.7). Nous distinguons deux cas pour lesquels nous considérons deux valeurs du pas de temps Δ : $\Delta \neq G_T$ et $\Delta = G_T$. Le temps de génération correspond ici à la période d'infectiosité, $G_T = \frac{1}{\gamma} = 2.5$ jours pour $\gamma = 0.4$. Ainsi, la valeur choisie pour Δ dans les deux cas est 0.5 et 2.5 jours respectivement. Comme l'illustrent Fine and Clarkson (1982), nous pouvons observer que les dynamiques de prévalence et d'incidence sont similaires lorsque le pas de temps est proche du temps de génération (Figure 1.7 droite). Lorsque le pas de temps Δ est assez différent du temps de génération ($\Delta = \frac{G_T}{5}$, Figure 1.7 gauche), les deux dynamiques sont très contrastées.

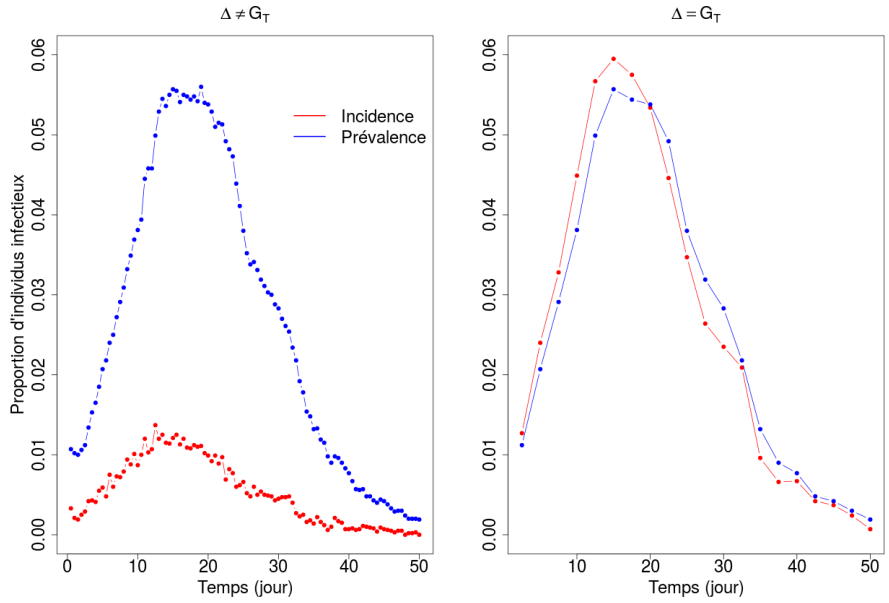

FIGURE 1.7 – Dynamiques épidémiques de la proportion de nouveaux individus infectieux (incidence, courbe rouge) et d'individus infectieux (prévalence, courbe bleue) modélisées par des processus Markoviens de sauts pour une taille de population $N = 10,000$ et pour différentes valeurs du pas de temps : $\Delta = 0.5 \neq G_T$ jours (graphique de gauche) et $\Delta = 2.5 = G_T$ jours (graphique de droite), où G_T est le temps de génération. Les valeurs des paramètres sont : $\lambda = 0.6$, $\gamma = 0.4$ et $\left(\frac{S(0)}{N}, \frac{I(0)}{N}, \frac{R(0)}{N}\right) = (0.95, 0.01, 0.04)$.

Exemple 17 : données d'incidence dans le modèle SIR Le nombre de nouveaux infectés au temps t_k correspond à

$$\int_{t_{k-1}}^{t_k} \lambda S(t) \frac{I(t)}{N}\, dt = S(t_{k-1}) - S(t_k),$$

et conditionnellement à $\Delta_k X = \Delta_k x$, il s'écrit sous la forme (1.13) avec $B = (-1 \quad 0)$.

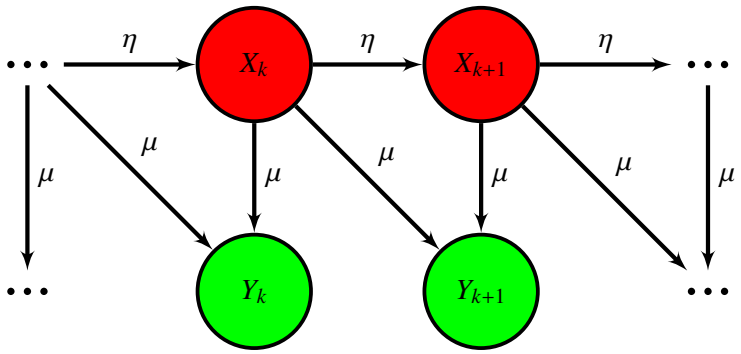

FIGURE 1.8 – Schématisation du modèle à espace d'état (1.15), où (X_k) est la variable stochastique (cercle) cachée (en rouge), (Y_k) est l'incidence observée (en vert), η est un vecteur de paramètres mécanistes et μ est un vecteur de paramètres d'observation.

1.2.2 Modélisation du bruit d'observation

En pratique, les données épidémiques sont sujettes à des erreurs de mesure et de report qu'il est important de prendre en compte dans la modélisation statistique. Pour cela, nous nous concentrons ici sur l'équation des observations des modèles à espace d'état (1.11) (resp. (1.15)) pour des données de prévalence (resp. incidence). Dans ces modèles, la prise en compte du bruit dans

les observations se fait à travers les distributions conditionnelles (1.11c) (prévalence) ou (1.15c) (incidence) et le vecteur des paramètres μ.

Différentes distributions des observations sont utilisées pour modéliser le bruit dans les données (voir e.g. Bretó (2018)). Le choix du modèle d'observations est très spécifique aux données disponibles et dépend de la façon dont elles sont recueillies et de la fiabilité des méthodes de report. Nous évoquons ici celles qui sont les plus fréquemment utilisées dans la littérature, avec une description de différents exemples d'application sur des données réelles. Sans perdre en généralité et par souci de clarté, nous considérons uniquement le cas de la prévalence pour décrire les différentes distributions.

— La distribution binomiale est classiquement utilisée pour modéliser des erreurs de report dans les données (Blackwood et al. (2013b), Blake et al. (2014)) :

$$Y_k|(X_k;\mu) \sim \text{Binomiale}\,(X_k, p)\,,\ \mu = p,$$

où p est un *taux de report*. Ainsi, les individus sont échantillonnés indépendamment avec une probabilité p.

Pour mieux comprendre la persistance de pathogènes dans des populations de chauve-souris vampires (*Desmodus rotundus*) en Amérique Latine, Blackwood et al. (2013b) ont considéré que la probabilité qu'une chauve-souris développe une infection fatale (*i.e.* observable) est égale à p. Pour estimer la contribution des transmissions de la poliomyélite chez les individus les plus âgés, Blake et al. (2014) ont utilisé des données de séries temporelles d'infections confirmées provenant de deux grandes épidémies ayant touché les individus adultes (Tadjikistan et République du Congo, 2010). Les auteurs ont considéré qu'un petit nombre seulement d'individus infectés ont été reportés et ont supposé que ce nombre suit une loi binomiale avec un taux de report p.

— La distribution de Poisson (Camacho et al. (2011), Shrestha et al. (2011)), naturelle pour décrire des comptages, est un deuxième exemple de distribution pour modéliser le processus d'observation :

$$Y_k|(X_k;\mu) \sim \text{Poisson}\,(pX_k)\,,\ \mu = p.$$

A la différence de la distribution binomiale, la distribution de Poisson n'est pas à support borné et suppose que les observations sont équidispersées.

Camacho et al. (2011) se sont intéressés à des données quotidiennes d'infection lors d'une double vague d'épidémie de grippe A/H3N2 survenant sur l'île Tristan da Cunha en 1971. Afin de modéliser le fait qu'une proportion égale à 85% du nombre total de cas avérés a été reportée et de prendre en compte d'éventuels cas asymptomatiques non reportés, les auteurs ont supposé que les observations suivent une loi de Poisson avec un taux de report p qu'ils ont ensuite estimé.

— Afin de prendre en compte une sur-dispersion dans les données, il est possible d'utiliser la distribution binomiale négative (voir e.g. Breto et al. (2009)) :

$$Y_k|(X_k;\mu) \sim \text{NegBin}\left(\frac{1}{\tau}, \frac{\tau p X_k}{1+\tau p X_k}\right),\ \mu = (p,\tau),$$

où $\tau > 0$ correspond à un *paramètre de sur-dispersion*. Ceci résulte d'une modélisation hiérarchique du processus de report. Soient ρ_k des taux de reports modélisés par des variables

aléatoires indépendantes suivant une loi Gamma$(1/\tau, p\tau)$. Conditionnellement à ρ_k, les observations sont modélisées comme des comptages de Poisson indépendants : $Y_k|(X_k, \rho_k) \sim$ Poisson$(\rho_k X_k)$. Alors, $Y_k|(X_k, \mu)$ suit une loi binomiale négative d'espérance $\mathbb{E}(Y_k|(X_k, \mu)) = pX_k$ et de variance $\text{Var}(Y_k|(X_k, \mu)) = pX_k + \tau p^2 X_k^2$.
Pour analyser des dynamiques de rougeole, Breto et al. (2009) ont considéré un modèle de type SEIR avec démographie. En notant $N_{IR}(t_k)$ le flux sortant du compartiment I et allant dans le compartiment R, où t_k est le k-ème temps d'observation à l'échelle de l'année, les cas de rougeole observés, agrégés de façon bihebdomadaire, sont notés $C_k = N_{IR}(t_k) - N_{IR}(t_{k-1})$. Conditionnellement à C_k, la distribution spécifiée pour les observations est une loi binomiale négative.

— Enfin, la distribution Gaussienne peut être choisie pour approximer les moyenne et variance d'une distribution donnée (voir e.g. Blackwood et al. (2013a), Lavine et al. (2013)) telle que la distribution binomiale

$$Y_k|(X_k; \mu) \sim \mathcal{N}\left(pX_k, p(1-p)X_k\right), \ \mu = p,$$

ou la distribution binomiale négative

$$Y_k|(X_k; \mu) \sim \mathcal{N}\left(pX_k, pX_k + \tau p^2 X_k^2\right), \ \mu = (p, \tau).$$

La distribution Gaussienne est à support continu et peut requérir une normalisation des données par la taille de population.
Blackwood et al. (2013a) ont analysé des données d'incidence mensuelle de cas de coqueluche en Thaïlande en considérant que les reports correspondent à une fraction p du nombre courant d'individus infectieux $I(t_k)$ aux temps t_k. Les auteurs ont supposé que le processus d'observation est normalement distribué de moyenne $pI(t_k)$ et de variance $\psi pI(t_k) + c$, où $\psi > 0$ est assimilé à une erreur de mesure et $c = 1$ est une constante additive incorporée dans le modèle pour prévenir d'éventuels écarts-type trop petits quand le nombre de cas $I(t_k)$ est très faible. Lavine et al. (2013) se sont intéressés à des séries temporelles de reports hebdomadaires de coqueluche provenant de la municipalité de Copenhague entre 1900 et 1937. Les auteurs ont considéré un modèle d'observation Gaussien, celui-ci approximant une loi binomiale sur-dispersée, d'espérance pC_k et de variance $p(1-p)C_k + \tau C_k^2$ avec C_k le nombre réel de cas. La sur-dispersion est proposée pour prendre en compte le fait que certains médecins ont tendance à reporter plus de cas durant le pic de l'épidémie (Metcalf et al. (2009)).

1.2.3 Modélisation simultanée de plusieurs dynamiques épidémiques

1.2.3.1 Contexte

Dans certaines situations, les observations disponibles proviennent de sites géographiques ou de périodes de temps distinctes et reflètent une variabilité géographique ou temporelle dans les mécanismes de propagation des épidémies. Pour prendre en compte cette variabilité inter-épidémie, une approche possible consiste à considérer les observations comme des *données longitudinales*, c'est-à-dire des observations répétées obtenues sur plusieurs *individus statistiques*. Dans notre contexte, les individus statistiques sont les différents sites géographiques ou les différentes périodes de temps qui structurent les données épidémiques. Dans certains domaines tels que l'économétrie, ce type de données est aussi connu sous le terme de *données de panel* (e.g. Hsiao (2014)), dénomination reprise par Bretó et al. (2020) en épidémiologie.

Un exemple parlant de données longitudinales en épidémiologie concerne les données épidémiques de la grippe humaine, celle-ci produisant une épidémie chaque année. Dans la Figure 1.9 sont représentées 28 dynamiques annuelles du nombre de syndromes grippaux sur 100,000 habitants en France entre 1990 et 2017, où chaque dynamique est recalée sur la même échelle de temps pour la visualisation. Le seuil d'incidence choisi pour définir la période épidémique est de 160 sur 100,000 habitants. Les données sont fournies par le réseau Sentinelles (url : `http://www.sentiweb.fr`). Nous pouvons observer des différences importantes entre les dynamiques, avec des pics épidémiques et des durées d'épidémie d'amplitude variable. Ces différences annuelles pourraient s'expliquer par la prédominance de différentes souches pour lesquelles les taux de transmission et de guérison seraient variables, ou encore par des changements dans le processus de report des données au fil des années. Dans de telles situations, un intérêt majeur consiste à analyser cette variabilité inter-épidémies afin de mieux comprendre les mécanismes de transmission récurrente et/ou multisite d'agents infectieux.

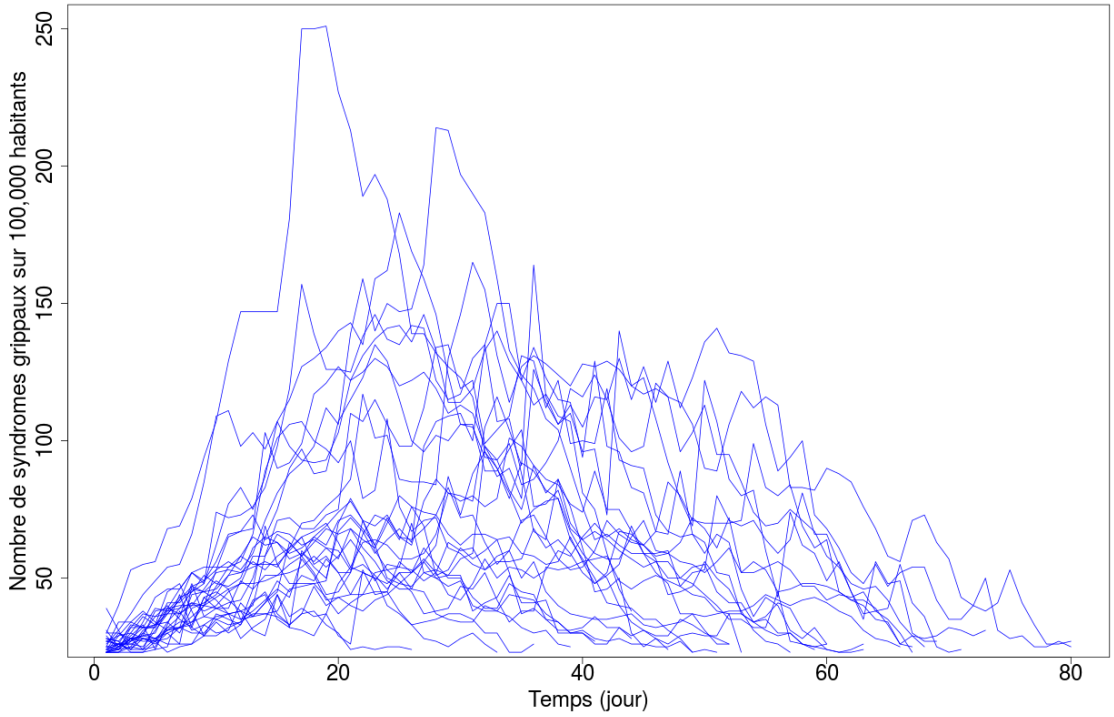

FIGURE 1.9 – 28 dynamiques annuelles, recalées sur la même échelle de temps, du nombre de syndromes grippaux sur 100,000 habitants en France entre 1990 et 2017 (réseau Sentinelles, url : `http://www.sentiweb.fr`).

Dans la suite, nous notons U le nombre d'individus statistiques et n_u le nombre d'observations par individu statistique u. De plus, $Y_u = (Y_{u,1}, \ldots, Y_{u,n_u})$, avec $Y_{u,k} := Y_u(t_{u,k})$, est le vecteur des données mesurées pour le u-ème individu statistique, $t_{u,k}$, $k = 1, \ldots, n_u$, sont les temps d'observation correspondants et $X_u = (X_{u,1}, \ldots, X_{u,n_u})$ sont les discrétisés des états du système.

Lorsque plusieurs dynamiques d'une même épidémie sont observées, il est pertinent d'utiliser le même modèle statistique pour décrire les observations relatives à chacune d'entre elles, mais avec des paramètres potentiellement différents d'une dynamique à l'autre pour tenir compte des différences existantes entre les dynamiques.

Il s'agit ici de considérer que pour tout $u = 1, \dots, U$:

$$\begin{cases} X_{u,k}|(X_{u,k-1};\eta_u) & \sim f(X_{u,k-1};\eta_u),\ k = 1,\dots,n_u, \quad (1.16a) \\ X_{u,0} & \sim f(X_{u,0};\eta_u), \quad (1.16b) \\ Y_{u,k}|(X_{u,k};\mu_u) & \sim g(X_{u,k};\mu_u),\ k = 0,\dots,n_u. \quad (1.16c) \end{cases}$$

Par analogie avec (1.11), η_u est le vecteur des paramètres régissant la u-ème dynamique et μ_u est le vecteur des paramètres du modèle reliant les observations de la u-ème dynamique à son processus caché. De plus, nous considérons que chaque dynamique provient d'un même processus mécaniste sous-jacent décrit par $f(\cdot)$ et leurs observations sont générées selon une même procédure d'observation décrite par $g(\cdot)$.

Dans la suite, nous noterons $\phi_u = (\eta_u, \mu_u)$ et utiliserons le terme de *paramètre individuel* pour désigner ce paramètre. La variabilité inter-individuelle (ou inter-épidémies) est alors décrite en termes de différences entre les ϕ_u, $u = 1, \dots, U$, dans la population. Deux approches sont possibles avec des implications différentes du point de vue de l'estimation.

1.2.3.2 Modèles à effets fixes

Une première possibilité est de considérer des effets individuels fixes. Cela revient à considérer autant de modèle de la forme (1.16) que de nombre U de paramètres individuels ϕ_u. Dans ce cas, en notant $c = \dim(\phi_u)$, le nombre total de paramètres à estimer est égal à $c \times U$.

C'est l'approche privilégiée par Bretó et al. (2020) pour analyser plusieurs dynamiques épidémiques simultanément mais aussi par Cazelles et al. (2021) afin d'étudier plusieurs dynamiques épidémiques de SARS-CoV-2 dans différentes régions de France et en Irlande, où chaque dynamique est analysée séparément selon un modèle SEIR étendu.

L'inconvénient de cette modélisation est que l'espace des paramètres peut être très élevé lorsque U est grand, car chaque vecteur des paramètres ϕ_u est estimé indépendamment des autres. Une autre conséquence est que la variabilité inter-individus est alors estimée empiriquement.

1.2.3.3 Modèles à effets aléatoires ou mixtes

Une deuxième possibilité consiste à décrire les différences entre individus statistiques en définissant les effets individuels comme des effets aléatoires. Cela conduit au cadre des modèles à effets mixtes, naturel pour modéliser une variabilité entre sujets au sein d'une même population à partir de données répétées (Lavielle (2014), Pinheiro and Bates (2000)), et particulièrement utilisé en pharmacocinétique. Dans ces modèles, les paramètres ϕ_u sont considérés comme des variables aléatoires indépendantes de distribution connue aux paramètres près. C'est la représentation que nous avons privilégiée dans cette thèse. Nous en détaillons ci-dessous le formalisme général.

Les modèles à effets mixtes peuvent être définis comme des modèles hiérarchiques à deux niveaux (cf Figure 1.10) :

— le premier niveau définit le *modèle individuel* et décrit la variabilité intrinsèque à chaque individu statistique via le modèle (1.16) :

$$\begin{cases} X_{u,k}|(X_{u,k-1};\eta_u) & \sim f(X_{u,k-1};\eta_u),\ k = 1,\dots,n_u, \\ X_{u,0} & \sim f(X_{u,0};\eta_u), \\ Y_{u,k}|(X_{u,k};\mu_u) & \sim g(X_{u,k};\mu_u),\ k = 0,\dots,n_u, \end{cases}$$

où $\phi_u = (\eta_u, \mu_u)$ avec $\phi_1, \ldots, \phi_U$ des vecteurs de paramètres individuels à valeurs dans $\mathbb{R}^c$;

— le deuxième niveau définit le *modèle de population* et décrit les différences entre individus statistiques. Les paramètres individuels $\phi_1, \ldots, \phi_U$ sont définis comme des variables aléatoires i.i.d. En général, ces variables sont supposées suivre une distribution Gaussienne (à une transformation près) :

$$h(\phi_u) = \alpha + \beta c_u + \xi_u,\ \xi_u \sim_{i.i.d} \mathcal{N}_c(0, \Gamma),\ u = 1, \ldots, U, \tag{1.17}$$

où $h(\phi) = (h_1(\phi_1), \ldots, h_c(\phi_c))' : \mathbb{R}^c \to \mathbb{R}^c$ est une fonction de lien connue, $\alpha \in \mathbb{R}^c$ est l'intercept, $c_u \in \mathbb{R}^p$ est un vecteur de covariables spécifiques à l'individu statistique u, β est une matrice de taille $c \times p$ d'effets fixes à valeurs réelles et $\xi_1, \ldots, \xi_U$ sont des effets aléatoires modélisés par U variables aléatoires centrées i.i.d tels que $\xi_u \in \mathbb{R}^c$.

Dans l'équation (1.17), la variabilité entre individus statistiques est séparée entre une variabilité attribuable à des covariables mesurées (c_u), la part restante étant décrite à travers les effets aléatoires (ξ_u). Les paramètres $\theta = (\alpha, \beta, \Gamma)$ sont généralement appelés paramètres de population, où α et β décrivent une tendance moyenne tandis que Γ quantifie la variabilité autour de cette tendance moyenne. Contrairement à l'approche précédente, ce ne sont plus les ϕ_u qui sont estimés mais les paramètres θ de leur distribution commune. La taille de l'espace des paramètres s'en trouve donc réduit. Ici, les ϕ_u sont des variables latentes.

A notre connaissance, les modèles à effets mixtes ont été peu utilisés pour décrire des dynamiques épidémiques, notamment dans le cas où les dynamiques latentes sont modélisées de façon stochastique. Par exemple, Prague et al. (2020) et Collin et al. (2021) ont analysé des données de SARS-CoV-2 dans plusieurs régions de France à l'aide d'un modèle SEIR étendu défini par un système d'EDO incorporant des paramètres aléatoires.

Remarque : Dans les modèles à effets fixes comme dans les modèles à effets mixtes présentés ci-dessus, les individus statistiques sont indépendants. Dans notre contexte, cela signifie que les dynamiques épidémiques évoluent indépendamment les unes des autres.

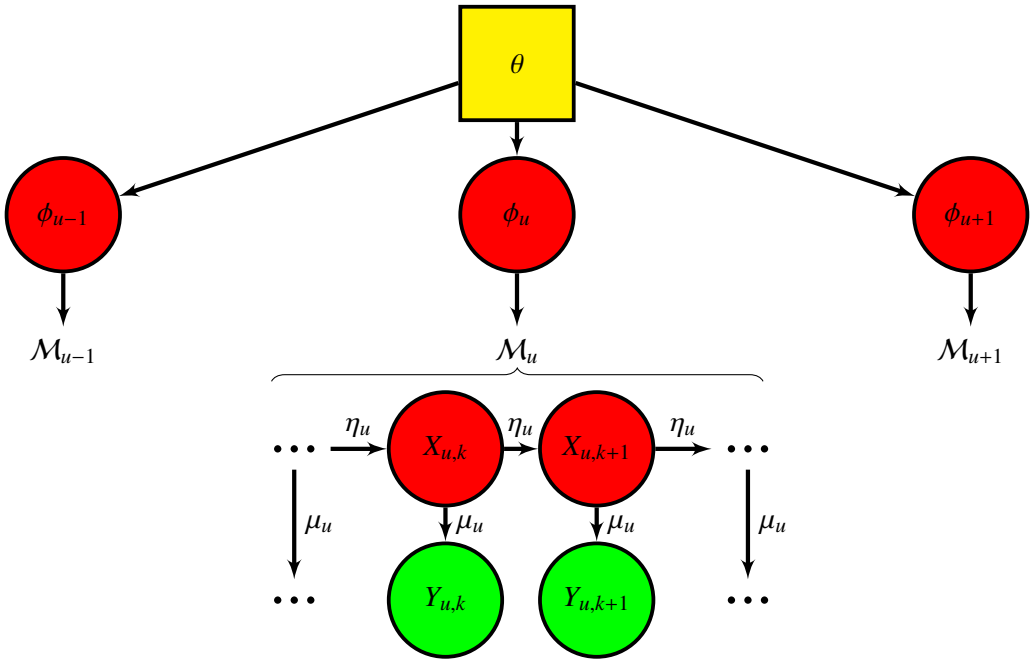

FIGURE 1.10 – Schématisation d'un modèle générique à effets mixtes à dynamique latente. Pour tout $u = 1, \ldots, U$, $(X_{u,k})_{k\geq 0}$ et $\phi_u = (\eta_u, \mu_u)$ sont des variables stochastiques (cercle) cachées (en rouge), $(Y_{u,k})_{k\geq 1}$ est la variable observée (en vert) et θ est un vecteur déterministe (carré) à estimer (en jaune). Chaque modèle $\mathcal{M}_u$ est un modèle à espace d'état de la forme (1.16).

1.3 Quelques algorithmes pour l'inférence

L'ensemble des modèles décrits ci-dessus sont des modèles à *données incomplètes*. En particulier, les dynamiques X_u et/ou, dans le cas de données longitudinales (ou répétées), les paramètres aléatoires ϕ_u ne sont pas observés. Dans ces modèles, l'inférence est généralement complexe du fait de la présence de variables latentes qui induisent, dans le calcul de la vraisemblance, des calculs d'intégrales souvent non explicites. De ce fait, pour obtenir des estimateurs de maximum de vraisemblance, des algorithmes sont nécessaires. Dans le Tableau 1.1, nous faisons une revue non exhaustive de ces algorithmes.

Par exemple, on parle d'approches *plug-and-play* les algorithmes qui spécifient le modèle dynamique, correspondant à la couche cachée, via un simulateur. Lorsque la simulation du modèle dynamique d'intérêt est simple, ce qui peut être le cas dans les modèles épidémiques, plusieurs algorithmes d'inférence sont développés. Dans le cadre des modèles à processus de Markov partiellement observés, ces algorithmes sont implémentés et réunis dans un même package R POMP (King et al. (2017)). D'autres approches s'intéressent plutôt au cadre Bayésien, par exemple les méthodes ABC (Marin et al. (2012)) fondées sur des statistiques résumées. Pour plus d'informations sur d'autres algorithmes d'inférence dans le cadre applicatif, le lecteur peut se référer à Britton and Giardina (2016) et King et al. (2017).

Dans cette section, nous détaillons certains algorithmes d'inférence ayant vocation à faire de l'estimation paramétrique dans des modèles à variables latentes par maximum de vraisemblance, sans être exhaustif. Plus précisément, nous distinguons dans la suite les algorithmes spécifiques des modèles à espace d'état de ceux pour le cadre plus général des modèles à variables latentes.

	Méthode	Référence
Fréquentiste	Iterated Filtering (MIF)	Ionides et al. (2006), Ionides et al. (2015)
	EM et ses variantes	Dempster et al. (1977), Wei and Tanner (1990), Delyon et al. (1999)
	Filtre de Kalman	Cappé et al. (2005)
Bayésien	PMCMC	Andrieu et al. (2010)
	ABC	Marin et al. (2012)

TABLE 1.1 – Différents algorithmes d'inférence applicables dans le cas des modèles à variables latentes et des modèles à espace d'état, classés selon leur appartenance au cadre fréquentiste ou Bayésien.

1.3.1 Algorithmes pour les modèles à espace d'état

Les modèles considérés dans cette section correspondent aux modèles à espace d'état de la forme (1.11). De façon générique, nous noterons n la taille de l'échantillon, $Y_{0:n} = (Y_0, \dots, Y_n)$ les variables observées, $X_{0:n} = (X_0, \dots, X_n)$ les variables latentes et $\theta = (\eta, \mu)$ les paramètres inconnus du modèle. On cherche à estimer θ par maximum de vraisemblance, c'est-à-dire évaluer :

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(Y_{0:n}; \theta),$$

où $L(Y_{0:n}; \theta)$ est la vraisemblance des observations $Y_{0:n}$ en les paramètres θ avec θ appartenant à un espace paramétrique Θ. Dans la suite, nous utilisons la notation $p(\cdot; \theta)$ pour désigner la fonction de densité d'une variable aléatoire quelconque en le paramètre θ.

Les algorithmes pour les modèles à espace d'état de la forme (1.11) sont souvent basés sur des méthodes de filtrage. En effet, les distributions de filtrage $p(X_k|Y_{0:k-1}; \theta)$, $Y_{0:k-1} = (Y_0, \dots, Y_{k-1})$, sont centrales dans l'expression de la vraisemblance :

$$L(Y_{0:n}; \theta) = p(Y_0; \theta) \prod_{k=1}^{n} p(Y_k|Y_{0:k-1}; \theta) = p(Y_0; \theta) \prod_{k=1}^{n} \int p(Y_k|X_k; \theta) p(X_k|Y_{0:k-1}; \theta)\, dX_k, \quad (1.18)$$

avec

$$p(X_k|Y_{0:k-1}; \theta) = p(X_k|X_{k-1}) p(X_{k-1}|Y_{0:k-1}).$$

Dans la suite, nous décrivons deux méthodes de filtrage connues : le filtre de Kalman et le filtre particulaire. Ces deux méthodes ont pour but de reconstruire la suite (X_k) des états cachés.

1.3.1.1 Filtre de Kalman

Les techniques de filtrage de Kalman s'appliquent à des modèles à espace d'état Gaussiens et linéaires en les états du système, qui peuvent être représentés comme suit (Cappé et al. (2005)) :

$$\begin{cases} X_k &= A_{k-1} X_{k-1} + R_{k-1} U_{k-1}, \ k \geq 1, \\ Y_{k-1} &= B_{k-1} X_{k-1} + S_{k-1} V_{k-1}, \end{cases} \quad (1.19)$$

où $(A_k)_{k\geq 0}$, $(B_k)_{k\geq 0}$, $(R_k)_{k\geq 0}$ et $(S_k)_{k\geq 0}$ sont connus et peuvent dépendre du paramètre θ (que nous omettons pour plus de lisibilité), et $(U_k)_{k\geq 0}$ et $(V_k)_{k\geq 0}$ sont deux séquences de variables aléatoires indépendantes entre elles et indépendantes des états (X_k), et distribuées selon une loi Gaussienne : $U_k \sim \mathcal{N}(0, I_d)$ et $V_k \sim \mathcal{N}(0, I_d)$, avec I_d la matrice identité. De plus, X_0 est supposée suivre une distribution Gaussienne $\mathcal{N}(\xi_0, R_0)$ et indépendante de (U_k) et (V_k).

Généralement, les techniques de filtrage de Kalman sont utilisées pour reconstruire les états cachés (X_k). Dans cette thèse, nous considérerons une autre approche dont l'originalité est d'utiliser les équations du filtre pour calculer les quantités impliquées dans le calcul de la vraisemblance. Nous en décrivons les aspects ci-dessous.

Dans (1.19), la vraisemblance des observations est explicite et son calcul requiert les espérances $\hat{X}_k$ et variances $\hat{\Xi}_k$ des distributions de filtrage $p(X_k|Y_{0:k-1};\theta)$. Les quantités $\hat{X}_k$ et $\hat{\Xi}_k$ sont alors calculables de façon exacte et récursive en utilisant les équations du filtre de Kalman ci-dessous. Sachant $\hat{X}_0 = \xi_0$, $\hat{\Xi}_0 = R_0$, pour $k \geq 0$:

$$\begin{aligned}
\hat{X}_k &= A_{k-1}\bar{X}_{k-1}, \\
\hat{\Xi}_k &= A_{k-1}\bar{R}_{k-1}A_{k-1}^t + R_{k-1}R_{k-1}^t, \\
\bar{X}_{k-1} &= \hat{X}_{k-1} + \hat{\Xi}_{k-1}B_{k-1}^t(B_{k-1}\hat{\Xi}_{k-1}B_{k-1}^t + S_{k-1}S_{k-1}^t)^{-1}(Y_{k-1} - B_{k-1}\hat{X}_{k-1}), \\
\bar{R}_{k-1} &= \hat{\Xi}_{k-1} - \hat{\Xi}_{k-1}B_{k-1}^t(B_{k-1}\hat{\Xi}_{k-1}B_{k-1}^t + S_{k-1}S_{k-1}^t)^{-1}B_{k-1}\hat{\Xi}_{k-1}.
\end{aligned}$$

A partir de ces quantités, il est alors possible d'avoir une expression explicite de (1.18) :

$$\log L(Y_{0:n};\theta) = C + \log p(Y_0;\theta) - \frac{1}{2}\sum_{k=1}^{n}\left[\log\left(|\hat{\Omega}_k|\right) + (Y_k - \hat{M}_k)^t\left(\hat{\Omega}_k\right)^{-1}(Y_k - \hat{M}_k)\right],$$

où

$$\begin{aligned}
\hat{M}_k &= B_k\hat{X}_k, \\
\hat{\Omega}_k &= B_k\hat{\Xi}_kB_k^t + S_kS_k^t,
\end{aligned}$$

C est une constante et $|A|$ désigne le déterminant de la matrice A.

Remarque : Lorsque le modèle (1.11) est Gaussien mais n'est plus linéaire en les états, il est possible d'utiliser la même approche sur une linéarisation du modèle suivant :

$$\begin{cases} X_k &= f(X_{k-1}, U_{k-1}),\ k \geq 1, \\ Y_{k-1} &= g(X_{k-1}, V_{k-1}). \end{cases}$$

Dans ce cas, le calcul de vraisemblance n'est plus exact mais approché. Le filtre est alors qualifié de filtre de Kalman étendu.

1.3.1.2 Méthodes particulaires

Lorsque les modèles ne sont pas linéaires ni Gaussiens, une alternative possible est le filtre particulaire, décrit par une procédure de type Sequential Monte Carlo (SMC, cf Doucet et al. (2001) pour plus de détails). Celui-ci reconstruit séquentiellement les variables d'états non-observées X_k pour obtenir une estimation de la distribution marginale $p(Y_k|Y_{0:k-1};\theta)$ puis une approximation de la vraisemblance des observations (1.18).

L'idée derrière les filtres particulaires est de générer un ensemble de particules qui approche la distribution $p(X_k|Y_{0:k-1};\theta)$ pour tout k. Pour cela, à chaque itération k de l'algorithme, l'état du système X_k est décrit par un nuage de particules $\left(X_k^{(j)}\right)_{1\leq j\leq J}$. Ces dernières se propagent alors à l'instant $k+1$ selon les modèles d'état et d'observation au prochain point d'observation (cf **Algorithme 1** pour le pseudo-code).

Algorithm 1: Sequential Monte Carlo (filtre particulaire)

Initialisation des particules : simuler $X_0^{(j)} \sim p(X_0; \theta)$ pour $j = 1 : J$;
for $k = 1 : n$ **do**
  Simuler $X_k^{(j)} \sim p(X_k | X_{k-1}^{(j)}; \theta)$ pour $j = 1 : J$;
  Calculer les poids $\omega_k^{(j)} = p(Y_k | X_k^{(j)}; \theta)$ pour $j = 1 : J$;
  Normaliser les poids $\tilde{\omega}_k^{(j)} = \omega_k^{(j)} / \sum_{m=1}^{J} \omega_k^{(m)}$ pour $j = 1 : J$;
  Ré-échantillonner les particules $X_k^{(j)}$ selon les poids $\tilde{\omega}_k^{(j)}$ pour $j = 1 : J$;
end

Il est possible de déduire une approximation empirique de la vraisemblance à partir des particules générées.

Par exemple, les algorithmes Particle Markov Chain Monte Carlo (PMCMC, Andrieu et al. (2010)) combinent une procédure d'évaluation de la vraisemblance $L(Y_{0:n}; \theta)$ à partir du filtre particulaire et un schéma d'estimation Markov Chain Monte Carlo (e.g. l'algorithme de Metropolis-Hastings). Ceux-ci ont été appliqués pour analyser par exemple des données épidémiques d'Ebola (Funk et al. (2018)), de peste bubonique (Didelot et al. (2017)) et de SARS-CoV-2 (Cazelles et al. (2021)). Le filtre particulaire est impliqué dans un autre algorithme d'inférence souvent utilisé en épidémiologie, l'algorithme Maximum Iterated Filtering (MIF, Ionides et al. (2006), Ionides et al. (2015)), que nous détaillons car il sera par la suite évoqué et utilisé (cf **Chapitre 2**).

1.3.1.3 Algorithme MIF

L'algorithme MIF repose sur des techniques de filtrage itératif pour faire de l'estimation de paramètres (cf **Algorithme 2** pour le pseudo-code). L'idée générale de cette méthode est d'augmenter la couche latente par le paramètre du modèle θ, celui-ci suivant une marche aléatoire $(\theta_k)_{k \geq 0}$ de moyenne et variance $\mathbb{E}[\theta_k | \theta_{k-1}] = \theta_{k-1}$, $\mathrm{Var}(\theta_k | \theta_{k-1}) = \sigma^2 \Sigma$, où σ est un scalaire et Σ est une matrice diagonale symétrique définie positive (choisie arbitrairement). Au fil des itérations, l'intensité σ de la marche aléatoire décroît vers zéro, ce qui permet d'obtenir une estimation de θ une fois la convergence atteinte. En pratique, pour une valeur donnée de la perturbation σ_0, à chaque itération m de l'algorithme, la perturbation décroît selon $\sigma_m = a^{m-1} \sigma_0$ où $0 < a < 1$ est un taux de *refroidissement*.

Algorithm 2: Iterated Filtering (Ionides et al. (2015))

Soient $0 < a < 1$ (taux de refroidissement) et σ_0 (perturbation initiale) donnés :
for $m = 1 : M$ **do**
  Poser $\sigma_m = a^{m-1} \sigma_0$;
  Initialisation des particules : simuler $\theta_{0,m}^{(j)} \sim p(\theta | \theta_{m-1}^{(j)}; \sigma_m)$ et $X_{0,m}^{(j)} \sim p(X_0; \theta_{0,m}^{(j)})$ pour $j = 1 : J$;
  for $k = 1 : n$ **do**
    Simuler $\theta_{k,m}^{(j)} \sim p(\theta | \theta_{k-1,m}^{(j)}; \sigma_m)$ et $X_{k,m}^{(j)} \sim p(X_k | X_{k-1,m}^{(j)}; \theta_m^{(j)})$ pour $j = 1 : J$;
    Calculer les poids $\omega_{k,m}^{(j)} = p(Y_k | X_{k,m}^{(j)}; \theta_{k,m}^{(j)})$ pour $j = 1 : J$;
    Normaliser les poids $\tilde{\omega}_{k,m}^{(j)} = \omega_{k,m}^{(j)} / \sum_{u=1}^{J} \omega_{k,m}^{(u)}$ pour $j = 1 : J$;
    Ré-échantillonner les particules $X_{k,m}^{(j)}$ et $\theta_{k,m}^{(j)}$ selon les poids $\tilde{\omega}_{k,m}^{(j)}$ pour $j = 1 : J$;
  end
  Fixer $\theta_m^{(j)} = \theta_{n,m}^{(j)}$ pour $j = 1 : J$;
end

Un filtre particulaire étant appliqué à chaque itération de l'algorithme MIF, le coût en temps de calcul peut être considérable. De plus, la méthode nécessite la calibration de nombreux paramètres de réglage (e.g. nombre de particules J, nombre d'itérations M) pouvant avoir une influence à la fois sur le temps de calcul mais aussi sur la qualité des estimations. Des propriétés théoriques de convergence de l'algorithme vers l'estimateur du maximum de vraisemblance ont été prouvées (Ionides et al. (2006), Ionides et al. (2011)) et une version améliorée, permettant de meilleurs résultats numériques dans l'inférence des paramètres, a été proposée (Ionides et al. (2015)). La méthode a souvent été utilisée pour faire de l'inférence de paramètres sur des données épidémiques (e.g rotavirus : Stocks et al. (2018) ; grippe humaine : Shrestha et al. (2013), Camacho et al. (2011), coqueluche : Lavine et al. (2013), Ebola : King et al. (2015)) via son implémentation dans le package POMP (King et al. (2017)).

Une extension de l'algorithme MIF, appelée Panel Iterated Filtering (PIF, Bretó et al. (2020)), a été proposée dans le cadre des données de panel (aussi implémentée dans le package POMP). En particulier, en modélisant des dynamiques épidémiques latentes par des processus Markoviens de sauts partiellement observés, les auteurs se sont intéressés à des données d'incidence de polio dans plusieurs états des Etats-Unis entre 1932 et 1953 et à des données de surveillance sur le VIH dans trois villes des Etats-Unis (San Francisco, Denver et Chicago) entre 1992 et 1995.

1.3.2 Algorithme Expectation-Maximisation et ses variantes

Les algorithmes Expectation-Maximisation (EM, Dempster et al. (1977)) et ses variantes Monte Carlo EM (MCEM, Wei and Tanner (1990)) et Stochastic Approximation EM (SAEM, Delyon et al. (1999)) sont utilisés pour l'estimation de paramètres dans le cadre des modèles à variables latentes, plus large que celui des modèles à espace d'état. Pour en décrire le fonctionnement, nous nous placerons dans un cadre général où le modèle, paramétré par θ, est composé d'une variable observée Y et d'une variable latente Z.

Dans ces modèles, la vraisemblance des observations $L(Y;\theta)$ s'exprime comme suit :

$$L(Y;\theta) = \int p(Y|Z;\theta)p(Z;\theta)\, dZ. \tag{1.20}$$

On cherche alors à évaluer :

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, L(Y;\theta).$$

L'expression (1.20) étant rarement explicite, les algorithmes de type EM se basent plutôt sur la vraisemblance des données complètes $p(Y,Z;\theta)$ pour obtenir une estimation du paramètre θ.

1.3.2.1 Algorithme EM

L'idée clée de l'algorithme EM (cf Algorithme 3 pour un pseudo-code) est de maximiser de façon itérative l'espérance conditionnelle de la log-vraisemblance des données complètes (Z,Y) sachant les observations Y par rapport au paramètre θ.

Algorithm 3: Expectation-Maximisation (Dempster et al. (1977))

A l'itération $m > 1$:

- **Étape E** : calculer l'espérance de la log-vraisemblance des données complètes sachant les observations en la valeur courante du paramètre θ_{m-1}

$$Q(\theta|\theta_{m-1}) = \mathbb{E}\left(\log p(Z, Y; \theta)|Y, \theta_{m-1}\right).$$

- **Étape M** : mettre à jour de l'estimateur en maximisant $Q(\theta|\theta_{m-1})$

$$\theta_m = \arg\max_\theta Q(\theta|\theta_{m-1}).$$

Sous des hypothèses de régularité sur le modèle, des propriétés théoriques de convergence de l'algorithme EM vers un point stationnaire de la log-vraisemblance des observations ont été établies (Dempster et al. (1977), Wu (1983)).

Cependant, dans certaines situations, l'étape E n'est pas traitable car $Q(\theta|\theta_{m-1})$ se calcule au moyen d'intégrales n'ayant pas toujours d'expressions analytiques simples. Pour résoudre ce problème, il est possible d'utiliser des versions stochastiques de l'algorithme EM dans lesquelles la quantité $Q(\cdot|\theta_{k-1})$ est approchée stochastiquement.

1.3.2.2 Algorithme MCEM

L'algorithme MCEM (Wei and Tanner (1990)) fournit une approximation par Monte-Carlo de la quantité $Q(\cdot|\theta_{k-1})$ de l'étape E de l'algorithme EM en générant à chaque itération un grand nombre de réalisations des variables latentes à partir de la distribution conditionnelle par rapport aux observations $p(\cdot|Y; \theta_{m-1})$ (cf Algorithme 4 pour un pseudo-code). Notons K_m le nombre de réalisations des données non-observées $Z_m = (Z_m(1), Z_m(2), \ldots, Z_m(K_m))$ obtenues à l'itération m de l'algorithme. La quantité $Q(\theta|\theta_{m-1})$ est alors approchée via un calcul de la moyenne empirique de la log-vraisemblance des données complètes.

Les propriétés théoriques sont étudiées dans Chan and Ledolter (1995) et Fort and Moulines (2003). En pratique, la convergence de l'algorithme est influencée par le nombre K_m de réalisations des données non-observées, cela pouvant occasionner un coût de calcul considérable.

Algorithm 4: Monte-Carlo EM (Wei and Tanner (1990))

A l'itération $m > 1$:

- **Étape S** : simuler K_m réalisations des variables aléatoires Z_m sous la distribution conditionnelle sachant les observations $p(\cdot|Y; \theta_{m-1})$ pour un paramètre courant θ_{m-1}.
- **Étape E modifiée** : approcher $Q(\theta|\theta_{m-1})$ selon

$$Q(\theta|\theta_{m-1}) \approx \frac{1}{K_m} \sum_{l=1}^{K_m} \log p(Z_m(l), Y; \theta).$$

- **Étape M** : mettre à jour de l'estimateur en maximisant $Q(\theta|\theta_{m-1})$

$$\theta_m = \arg\max_\theta Q(\theta|\theta_{m-1}).$$

1.3.2.3 Approximation stochastique de l'EM

Algorithme SAEM

L'algorithme SAEM (Delyon et al. (1999) ; cf Algorithme 5 pour un pseudo-code), variante stochastique de l'algorithme EM, combine à chaque itération m la simulation des données non-observées sous la distribution conditionnelle sachant les observations $p(\cdot|Y;\theta_{m-1})$ en le paramètre courant θ_{m-1} (étape S) et une approximation stochastique de $Q(\theta|\theta_m)$ (étape SA).

Algorithm 5: Stochastic Approximation EM (Delyon et al. (1999))

A l'itération $m > 1$:

- **Étape S** : simuler une réalisation des variables aléatoires Z_m sous la distribution conditionnelle sachant les observations $p(\cdot|Y;\theta_{m-1})$ pour un paramètre courant θ_{m-1}.
- **Étape SA** : mettre à jour $Q_m(\theta)$ selon

$$Q_m(\theta) = Q_{m-1}(\theta) + \alpha_m(\log p(Z_m, Y;\theta) - Q_{m-1}(\theta)),$$

  où $(\alpha_m)_{m\geq 1}$ est une séquence de pas positifs tels que $\sum_{m=1}^{\infty}\alpha_m = \infty$ et $\sum_{m=1}^{\infty}\alpha_m^2 < \infty$.
- **Étape M** : mettre à jour l'estimateur du paramètre en maximisant $Q_m(\theta)$

$$\theta_m = \arg\max_\theta Q_m(\theta).$$

L'implémentation de l'algorithme SAEM peut être simplifiée dans le cas où la vraisemblance des données complètes appartient à une famille de modèles exponentiels courbe, *i.e.*

$$p(Y,Z;\theta) = \exp\left(-\psi(\theta) + \langle S(Z,Y), \varphi(\theta)\rangle\right), \tag{1.21}$$

avec $\psi(\cdot)$ et $\varphi(\cdot)$ des fonctions du paramètre θ, $S(\cdot)$ des statistiques exhaustives du modèle des données complètes (Z,Y) et $\langle\cdot,\cdot\rangle$ le produit scalaire. Dans ce cas, la mise en oeuvre des étapes SA et M est simplifiée (cf Algorithme 6 pour un pseudo-code).

Algorithm 6: Stochastic Approximation EM (famille exponentielle courbe)

A l'itération $m > 1$:

- **Étape S** : simuler une réalisation des paramètres aléatoires Z_m sous la distribution conditionnelle sachant les observations $p(\cdot|Y;\theta_{m-1})$ pour un paramètre courant θ_{m-1}.
- **Étape SA** : mettre à jour s_m selon

$$s_m = s_{m-1} + \alpha_m(S(Z_m, Y) - s_{m-1}),$$

  où $(\alpha_m)_{m\geq 1}$ est une séquence de pas positifs tels que $\sum_{m=1}^{\infty}\alpha_m = \infty$ et $\sum_{m=1}^{\infty}\alpha_m^2 < \infty$.
- **Étape M** : mettre à jour l'estimateur du paramètre selon

$$\theta_m = \arg\max_\theta \ (-\psi(\theta) + \langle s_m, \varphi(\theta)\rangle).$$

Lorsque la vraisemblance des données complètes peut s'écrire sous la forme (1.21), Delyon et al. (1999) ont montré que l'algorithme SAEM présenté ci-dessus converge vers un point stationnaire de la vraisemblance des observations. Du point de vue de l'implémentation, l'algorithme SAEM ne nécessitant qu'une seule simulation des données non-observées par itération (étape S), le coût de

calcul est considérablement plus faible que dans le cas de l'algorithme MCEM. Un autre avantage est, qu'en pratique, peu d'itérations sont nécessaires pour que l'algorithme converge.

Algorithme MCMC-SAEM

Dans certaines situations, un échantillonnage exact sous la distribution $p(\cdot|Y;\theta_{m-1})$ à l'étape S est impossible. Kuhn and Lavielle (2004) ont proposé de combiner l'algorithme SAEM avec des méthodes MCMC telles que l'algorithme Metropolis-Hastings (cf Algorithme 7 pour un pseudo-code).

Algorithm 7: Metropolis-Hastings

Pour une valeur de paramètre θ donnée, à l'itération $m > 1$:

- Générer un candidat $Z^{(c)}$ selon une loi de proposition $q(Z_{m-1},\cdot)$.
- Fixer

$$Z_m = \begin{cases} Z_{m-1} \text{ avec probabilité } 1-\alpha(Z_{m-1},Z^{(c)}), \\ Z^{(c)} \text{ avec probabilité } \alpha(Z_{m-1},Z^{(c)}), \end{cases}$$

où $\alpha(Z_{m-1},Z^{(c)})$ est le ratio d'acceptation s'écrivant

$$\alpha(Z_{m-1},Z^{(c)}) = \min\left[1, \frac{p(Y|Z^{(c)};\theta)\ p(Z^{(c)};\theta)\ q(Z_{m-1},Z^{(c)})}{p(Y|Z_{m-1};\theta)\ p(Z_{m-1};\theta)\ q(Z^{(c)},Z_{m-1})}\right].$$

Les propriétés de convergence sont étudiées dans Kuhn and Lavielle (2004). En pratique, un nombre faible d'itérations de la procédure MCMC dans l'algorithme SAEM suffisent (Kuhn and Lavielle (2005)).

Utilisation de l'algorithme MCMC-SAEM dans les modèles mixtes à dynamique latente

L'algorithme SAEM a souvent été utilisé pour faire de l'inférence de paramètres dans les modèles à effets mixtes (Lavielle (2014), Kuhn and Lavielle (2005)). Lorsque le modèle mixte est défini au moyen d'un processus dynamique latent, il existe deux catégories de variables cachées : les effets aléatoires (e.g. $(\phi_u)_{u\geq 1}$, cf (1.17)) et le processus dynamique caché (e.g. $(X_u)_{u\geq 1}$, cf (1.16)). Dans la littérature, le processus caché est souvent défini par une EDO (e.g. Donnet and Samson (2007)) ou une équation différentielle stochastique (EDS, e.g. Donnet and Samson (2008)). Lorsque la dynamique latente est modélisée par une EDS, la difficulté principale consiste à réaliser l'étape S de l'algorithme SAEM et en particulier, la simulation de la dynamique latente. Une première stratégie est d'approcher le processus par un schéma de discrétisation (e.g. Euler Maruyama, Donnet and Samson (2008)) et simuler la dynamique cachée selon ce schéma. Une autre stratégie, privilégiée par Donnet and Samson (2014), est d'utiliser des méthodes particulaires pour simuler la dynamique cachée. D'une façon alternative, Delattre and Lavielle (2013) ont proposé de marginaliser par rapport à la dynamique cachée à l'aide d'un filtre de Kalman étendu (cf Section 1.3.1.1) pour ne simuler que les effets aléatoires.

Les exemples d'application de l'algorithme SAEM à des modèles mixtes à dynamique latente sont nombreux en pharmacocinétique (nous référons le lecteur à Donnet and Samson (2013) pour une revue de plusieurs méthodes d'inférence applicables dans des modèles définis par des équations différentielles stochastiques). Cependant, à notre connaissance, l'approche a été peu utilisée pour inférer des paramètres gouvernant des dynamiques épidémiques. Prague et al. (2020) ont proposé

une stratégie d'estimation basée sur des techniques de filtrage de Kalman et l'utilisation de l'algorithme SAEM pour estimer les paramètres du modèle, mais dans le cas où les dynamiques latentes sont modélisées par un système d'EDO. A notre connaissance, une approche combinant l'utilisation de ces outils et une modélisation stochastique des dynamiques épidémiques n'a jamais été proposée.

1.4 Objectifs et contributions de la thèse

Dans cette thèse, nous avons adopté une modélisation stochastique des épidémies. En définissant un modèle pour les observations selon leurs nombreuses spécificités (cf points (i), (ii) et (iii) du **Préambule**), nous avons développé des algorithmes d'estimation paramétrique simples et performants. Nous en avons étudié les performances sur des jeux de données simulées, puis nous les avons appliqués à des données épidémiques réelles.

Le **Chapitre 2** a fait l'objet d'un article publié dans le journal Computational Statistics and Data Analysis (cf Narci et al. (2021a)) et intitulé :
"*Inference for partially observed epidemic dynamics guided by Kalman filtering techniques*".
Dans celui-ci, l'objectif est de proposer une méthode générique d'inférence facile à utiliser, avec peu de paramètres de réglage algorithmique à calibrer, et qui peut être appliquée à tout modèle à compartiments et à partir de données de prévalence bruitées.
Pour décrire les dynamiques épidémiques d'un système compartimental de taille d en cas d'épidémies majeures, nous considérons un processus Gaussien à temps continu $G_N(\eta, t)$, dépendant d'un vecteur de paramètres mécanistes η incluant les conditions initiales x_0, qui approxime le processus Markovien de sauts densité-dépendant $\mathcal{Z}_N(\cdot)$ défini en (1.3). En particulier, il satisfait :

$$G_N(\eta, t) = x(\eta, t) + \frac{1}{\sqrt{N}} g(\eta, t),$$

où $x(\cdot)$ est la solution déterministe du système (1.1), $g(\cdot)$ est un processus Gaussien non-homogène dans le temps tel que

$$g(\eta, t) = \int_0^t \Phi(t, s, \eta)\sigma(\eta, x(s, \eta))\, dB(s),$$

avec $(B(t))_{t\geq 0}$ un mouvement Brownien de dimension d, $\sigma(\cdot)$ la décomposition de Cholesky de la matrice de diffusion définie en (1.4) et $\Phi(\cdot)$ la matrice de taille $d \times d$ telle que

$$\Phi(\eta, t, s) = \exp\left(\int_s^t \nabla_x b(\eta, x(\eta, u))\, du\right),\ \nabla_x b(\eta, x) = \frac{\partial b_i}{\partial x_j}(\eta, x))_{1\leq i,j\leq d}.$$

Nous utilisons le formalisme des modèles à espace d'état dans l'objectif de poser un cadre statistique exploitable. Pour ce faire, nous proposons une écriture autorégressive de l'équation décrivant les états, $(X_k)_{k\geq 0}$, obtenue par une discrétisation du processus $G_N(\cdot)$ à des instants d'observation $t_k = k\Delta$, avec Δ un pas de temps, et une approximation Gaussienne du modèle des observations. Cela permet d'obtenir un modèle à espace d'état Gaussien et linéaire de la forme

$$\begin{cases} X_k &= F_k(\eta, \Delta) + A_{k-1}(\eta, \Delta)X_{k-1} + R_k(\eta, \Delta)U_k, \\ Y_k &= B(\mu)X_k + S_k(\theta)V_k, \end{cases} \tag{1.22}$$

où

$$\begin{aligned} A_{k-1}(\eta) &= A(\eta, t_{k-1}) = \Phi(\eta, t_k, t_{k-1}), \\ F_k(\eta) &= F(\eta, t_k) = x(\eta, t_k) - \Phi(\eta, t_k, t_{k-1})x(\eta, t_{k-1}), \end{aligned}$$

$(U_k)_{k\geq 0}$ et $(V_k)_{k\geq 0}$ sont des variables aléatoires Gaussiennes i.i.d et centrées, $B(\cdot)$ est la composée d'un opérateur de projection (cf Section 1.2.1) et d'un opérateur prenant en compte les erreurs d'observation, $R_k(\cdot)$ et $S_k(\mu, \cdot)$ sont des quantités dépendant des paramètres η et du pas de temps Δ, et $\theta = (\eta, \mu)$ avec μ un vecteur de paramètres du modèle des observations. Les différences avec les modèles à espace d'état classiques de la forme (1.19) sont qu'un terme de centrage est ajouté

dans l'équation des états et que les quantités $A_{k-1}(\cdot)$ et $F_k(\cdot)$ sont exprimées non-linéairement en les paramètres mécanistes et les conditions initiales.
L'objectif est d'estimer θ. A partir du modèle discrétisé (1.22), nous proposons une approche d'estimation paramétrique basée sur un contraste construit à partir de la log-vraisemblance des observations du modèle approché (cf Guy et al. (2015) qui ont montré que les estimateurs associés ont de bonnes propriétés théoriques). Ensuite, nous utilisons les équations du filtre de Kalman, non pas dans le but de reconstruire la dynamique cachée du modèle comme c'est généralement le cas, mais pour calculer les quantités impliquées dans l'expression de la log-vraisemblance des observations (cf Section 1.3.1.1).
Les performances de la méthode son évaluées sur des dynamiques SIR simulées selon des processus Markoviens de sauts et comparées à l'algorithme d'inférence Maximum Iterated Filtering en fonction de différents jeux de valeurs des paramètres, de la taille de population et du nombre d'observations. Les résultats obtenus montrent que les biais et la précision des estimations sont tout à fait satisfaisants et similaires entre les deux méthodes. En plus des données simulées, notre méthode est appliquée sur des données réelles d'épidémie de grippe dans un pensionnat Anglais en 1978. Les valeurs des paramètres estimés obtenues sont plausibles. Cela est en particulier conforté par le fait que les trajectoires épidémiques de prédiction, obtenues par simulation à partir du modèle épidémique avec les valeurs estimées des paramètres, sont consistantes avec les données.
Une difficulté dans l'implémentation de notre approche réside en particulier dans la dépendance non-linéaire par rapport aux paramètres et aux conditions initiales des quantités apparaissant dans les équations du filtre. Une originalité est que ce formalisme n'est pas rattaché à un pas de temps donné pour les observations et peut tout à fait inclure des pas de temps variables. De plus, un avantage considérable de notre méthode est qu'elle ne nécessite que très peu de paramètres de réglage algorithmiques, qu'elle est relativement peu sensible à l'initialisation des paramètres dans la procédure d'inférence et qu'elle ne nécessite pas de simulations d'états cachés. Cette dernière spécificité permet en particulier une réduction considérable du temps de calcul quand la taille de population est grande par rapport à des approches faisant appel à de nombreuses simulations du modèle épidémique afin de reconstruire la chaîne (X_k) du processus caché.

Le **Chapitre 3** a fait l'objet d'un article soumis (pre-print disponible, Narci et al. (2021b)) et intitulé :
"*Inference in Gaussian state-space models with mixed effects for multiple epidemic dynamics*".
Deux contributions principales sont apportées dans cet article.
Une première contribution est d'étendre la méthode présentée dans le **Chapitre 2** afin qu'elle soit applicable sur des données d'incidence bruitées. Pour cela, nous proposons un modèle à espace d'état Gaussien et linéaire dans lequel les états cachés sont les accroissements $\Delta_k X$ de X_k :

$$\begin{cases} \Delta_k X = G_k(\eta, \Delta) + (A_{k-1}(\eta, \Delta) - I_d) \sum_{l=1}^{k-1} \Delta_l X + R_k(\eta, \Delta) U_k, \\ Y_k = \tilde{B}(\mu) \Delta_k X + \tilde{S}_k(\theta) V_k, \end{cases} \tag{1.23}$$

où

$$G_k(\eta, \Delta) = x(\eta, t_k) - x_0 - \Phi(\eta, t_k, t_{k-1})(x(\eta, t_{k-1}) - x_0),$$

$\tilde{S}_k(\cdot)$ dépend des paramètres η et μ et du pas de temps Δ, et $\tilde{B}(\cdot)$ est la composée d'un opérateur de projection (cf Section 1.2.1) et d'un opérateur prenant en compte les erreurs d'observation. La séquence $(\Delta_k X)_{k \geq 1}$ dépendant de toute l'information passée, elle n'est pas Markovienne et ne possède donc pas les propriétés requises pour l'application de techniques classiques de filtrage de Kalman. Nous élaborons une procédure itérative et définissons un nouveau filtre permettant de calculer récursivement la log-vraisemblance des observations du modèle (1.23).
Une deuxième contribution est de généraliser l'approche présentée dans le **Chapitre 2**. L'objectif est de tenir compte des différentes sources de variabilité dans les données d'observations répétées

dans le temps (e.g. saison) ou dans l'espace (e.g. région) d'une même épidémie dans un cadre statistique unifié décrit par les modèles à effets mixtes. Pour cela, nous proposons une modélisation hiérarchique à deux niveaux : le premier niveau modélise la variabilité intra-épidémie par un modèle à espace d'état Gaussien et linéaire (e.g. (1.22) ou (1.23)) tandis que le deuxième niveau modélise la variabilité inter-épidémies en spécifiant une distribution pour les paramètres aléatoires ϕ_u du modèle (cf (1.17)). Par exemple, conditionnellement à $\phi_u = \varphi$, cela revient à étendre le modèle (1.22) comme suit :

$$\begin{cases} X_{u,k} = F_k(\varphi, \Delta) + A_{k-1}(\varphi, \Delta)X_{u,k-1} + R_k(\varphi, \Delta)U_k, \\ Y_{u,k} = B(\varphi)X_{u,k} + S_k(\varphi)V_k, \end{cases}$$

où l'indice u désigne ici une épidémie. De plus, nous considérons que les paramètres spécifiques à chaque épidémie $(\phi_u)_{u\geq 1}$ sont des variables aléatoires i.i.d telles que

$$\begin{cases} \phi_u &= h(\beta, \xi_u), \\ \xi_u &\sim \mathcal{N}(0, \Gamma), \end{cases}$$

avec $h(\cdot)$ une fonction de lien connue, β un vecteur d'effets fixes, $(\xi_u)_{u\geq 1}$ des effets aléatoires modélisés par des variables aléatoires centrées, i.i.d et de matrice de variance Γ (cf Section 1.2.3.3). Pour estimer les paramètres de population $\theta = (\beta, \Gamma)$, nous combinons l'algorithme stochastique MCMC-SAEM (cf Section 1.3.2.3) avec les techniques de filtrage de Kalman construites dans le **Chapitre 2** et utilisées pour marginaliser par rapport à la dynamique cachée $(X_{u,k})_{k\geq 0}$. Ici, l'intérêt est qu'il suffit de simuler les paramètres aléatoires non-observés (ϕ_u) dans l'étape S de l'algorithme SAEM.
Nous avons tout d'abord évalué sur des données simulées l'approche d'inférence utilisant de façon conjointe les dynamiques épidémiques avec une autre approche les considérant indépendamment les unes des autres. Ensuite, notre méthode d'inférence est appliquée à des données d'incidence de syndromes grippaux en France entre 1990 et 2017 fournies par le réseau Sentinelles. Les résultats obtenus ont montré que, par la prise en compte simultanée de toutes les dynamiques épidémiques dans une procédure d'inférence, la méthode à effets mixtes permet d'obtenir des estimations moins biaisées et une évaluation quantitative plus précise de la variabilité inter-épidémies (e.g. variabilité saisonnière de la grippe humaine).
Une contribution est ici d'avoir utilisé le cadre des modèles à effets mixtes dans lequel les dynamiques épidémiques latentes sont modélisées par des processus stochastiques. A notre connaissance, aucune étude en épidémiologie n'a considéré en détail ce cadre particulier. Ainsi, les variabilités inter- et intra-épidémies sont décrites dans un modèle unifié permettant une amélioration de l'inférence statistique comparée à des approches empiriques qui considèrent de façon indépendante les dynamiques épidémiques.

Le **Chapitre 4** comporte une étude sur les données d'incidence d'infections, d'hospitalisations et de décès de la Covid-19 dans plusieurs régions de France. Le modèle à compartiments choisi pour décrire de façon mécaniste les dynamiques épidémiques est un modèle SEIR étendu, prenant en compte les individus asymptomatiques (A), hospitalisés (H) et décédés (D) (modèle SEIRAHD). Les données de la Covid-19 montrent clairement que les dynamiques épidémiques sont variables selon la région. Cela peut être dû au processus de report spécifique à la région et/ou à la dynamique épidémique elle-même. Dans le but d'apporter une évaluation quantitative de la variabilité inter-épidémies et de la tendance moyenne à laquelle répondent les dynamiques épidémiques à l'échelle de la région, nous proposons un modèle à espace d'état Gaussien et linéaire (dont la forme générale est décrite dans le **Chapitre 3**) avec des effets mixtes sur les paramètres pour modéliser les dynamiques épidémiques. La procédure d'inférence combinant l'algorithme MCMC-SAEM avec des

techniques de filtrage de Kalman est alors appliquée sur ce modèle, d'abord sur données simulées puis sur données réelles, afin d'estimer les paramètres de population d'intérêt.
Une contribution de ce chapitre est d'avoir pris en compte dans un modèle unique les variabilités intra- et inter-épidémies. Ceci nous permet de proposer une estimation plus précise et rigoureuse des paramètres du modèle et de la variabilité inter-épidémies de la Covid-19 à l'échelle régionale.

Le manuscrit se termine par le **Chapitre 5** contenant une discussion et quelques perspectives.

Chapitre 2

Premier article : Inference for partially observed epidemic dynamics guided by Kalman filtering techniques.

Table des matières

2.1 Introduction **43**
2.2 Gaussian model approximation for large population epidemics **45**
2.2.1 Preliminary comments on inference in epidemic models 45
2.2.2 Approximation of large population epidemic models and the autoregressive point of view 45
2.2.3 Approximation of the observation model 48
2.2.4 Application on the SIR epidemic model 49
2.3 Parameter estimation using Kalman filtering techniques **51**
2.3.1 Approximate likelihood inference 51
2.3.2 Application on the SIR epidemic model 53
2.4 Simulation study **53**
2.4.1 Simulation settings 53
2.4.2 Inference : settings, performance comparison, and implementation . . . 54
2.4.3 Point estimates and standard deviations for key model parameters θ . . 56
2.4.4 Confidence interval estimates based on profile likelihood 61
2.5 Application on real data **62**
2.6 Discussion **63**

Note Ce chapitre a fait l'objet d'un article publié dans le journal Computational Statistics and Data Analysis, volume 164 (cf Narci et al. (2021a)).

Abstract Despite the recent development of methods dealing with partially observed epidemic dynamics (unobserved model coordinates, discrete and noisy outbreak data), limitations remain in practice, mainly related to the quantity of augmented data and calibration of numerous tuning parameters. In particular, as coordinates of dynamic epidemic models are coupled, the presence of unobserved coordinates leads to a statistically difficult problem. The aim is to propose an easy-to-use and general inference method that is able to tackle these issues. First, using the properties of epidemics in large populations, a two-layer model is constructed. Via a diffusion-based approach, a Gaussian approximation of the epidemic density-dependent Markovian jump process is obtained, representing the state model. The observational model, consisting of noisy observations of certain model coordinates, is approximated by Gaussian distributions. Then, an inference method based on an approximate likelihood using Kalman filtering recursion is developed to estimate parameters of both the state and observational models. The performance of estimators of key model parameters is assessed on simulated data of SIR epidemic dynamics for different scenarios with respect to the population size and the number of observations. This performance is compared with that obtained using the well-known maximum iterated filtering method. Finally, the inference method is applied to a real data set on an influenza outbreak in a British boarding school in 1978.

Keywords Approximate maximum likelihood ; Diffusion approach ; Kalman filter ; Measurement errors ; Partially-observed Markov process ; Epidemic dynamics.

2.1 Introduction

The interest and impact of mathematical modeling and inference methods for infectious diseases have considerably grown in recent years in a context of increasing complex models and abundant data of varying quality. Estimating the parameters governing epidemic dynamics from available data has become a major challenge, in particular from the perspective of subsequently providing reliable predictions of such dynamics.

Many authors have addressed the problem of key epidemic parameter estimation based on likelihood approaches (e.g., Cauchemez and Ferguson (2008)). While estimation is quite straightforward for complete observations, this is no longer true in the incomplete observation setting which occurs in practice, regardless of the mathematical formalism used. Indeed, available data tends to be only partially observed (e.g., certain health statuses such as asymptomatic infected stages cannot be observed at all ; infectious and recovery dates are not observed for all individuals during the outbreak ; not all infectious individuals are reported) and may also be temporally and/or spatially aggregated. Various approaches have been developed to deal with these types of data (e.g., see O'Neill (2010), Britton and Giardina (2016) for reviews). In the general framework of partially-observed Markov processes, some of these methods have been implemented in the R package POMP (King et al. (2017)). Among these, we cite maximum iterated filtering (MIF : Ionides et al. (2006), Ionides et al. (2015)) in which the parameter space is explored by considering that parameters follow a random walk over time with variance decreasing over filtering iterations, and the likelihood being stochastically estimated. Theoretical justification for convergence to the maximum likelihood estimates in the parameter space has been provided for this method (Ionides et al. (2011)). Furthermore, likelihood-free methods, such as approximate Bayesian computation based on sequential Monte Carlo (ABC-SMC, Sisson et al. (2007), Toni et al. (2009)) and particle Markov chain Monte Carlo (PMCMC, Andrieu et al. (2010)), have opened some of the most promising pathways for improvement. Nevertheless, these algorithms do not provide a definitive solution to statistical inference from incomplete epidemic data. Indeed, there are real limitations in practice due to the amount of augmented data and fitting the numerous tuning parameters invol-

ved. That can lead to substantial computational overheads.

In this article, we consider a different approach to deal with the presence of missing coordinates, discrete observations, and reporting and measurement errors. Our goal is to propose a useful and coherent latent variable model that allows key epidemic parameters to be estimated from imperfect observations from outbreaks.

A multidimensional Markov jump process describes the epidemic dynamics in a closed population of size N. Using the large population framework, i.e., with N large, we first build an approximation of epidemic dynamics using an autoregressive Gaussian process via a diffusion approach (see e.g., Ethier and Kurtz (2005), Guy et al. (2015)). Then we simultaneously account for a given missing coordinate value and systematic noise present in observations by applying a projection operator to the process and adding heteroscedastic Gaussian errors. This yields the theoretical framework that allows recursive computations of an approximate likelihood. This approach, based on Kalman filtering, enables the computation of the approximate log-likelihood of the available observations and, consequently, the estimation of model parameters. An initial innovative aspect of this method with respect to others is the use of a Kalman filter to recursively compute the approximate likelihood in the non-standard case of the small noise framework (i.e., with noise covariance matrix proportional to $1/N$), rather than the classical recurrent case coupled with a large observation time-window (with the number of observations going to infinity). In addition, the explicit integration into the algorithm of the data sampling interval, and an alternative point of view in the prediction of successive model states—given the observations—are further innovative points.

The derivation and accuracy assessment of Gaussian process approximation for stochastic epidemic models have previously been described in Buckingham-Jeffery et al. (2018), along with maximum likelihood inference for parameters underlying epidemic dynamics. However, that study does not rely on Kalman filtering, nor does it consider noise in outbreak data. Computation of the approximate likelihood of the associated statistical model, as well as parameter estimation, performed via Kalman filtering recursion was proposed in Favetto and Samson (2010), but for simpler models without nonlinear terms in the drift, and with no parameter to estimate in the diffusion term.

For the sake of simplicity, we consider here an epidemic with homogeneous mixing in a closed population whose dynamics are described by a compartmental model, with each compartment containing individuals with identical health states. We focus on the simple SIR (susceptible - infectious - recovered) epidemic model characterized by a two-dimensional jump process, partially observed at regularly-spaced discrete times, with measurement errors. The approach can be easily extended to broader epidemic models observed with various sampling intervals.

The article is organized as follows. In Section 2.2 we introduce the general framework and related inference issues, and propose the model approximation. Section 2.3 contains the main methodological developments of our paper : construction of the approximate log-likelihood, its computation based on Kalman filtering recursion, and the associated parameter estimation. In Sections 2.4 and 2.5 we assess the performance of our estimators on both simulated data and real data from an influenza outbreak in a British boarding school in 1978, and compare our results with those obtained using the MIF method. Section 2.6 contains a discussion and concluding remarks.

2.2 Gaussian model approximation for large population epidemics

2.2.1 Preliminary comments on inference in epidemic models

Epidemic dynamics can be naturally described using compartmental models, which are by essence mechanistic and include parameters in their characterization. In such models, the population is partitioned into compartments corresponding to different stages of the infection process, whose temporal evolution is described. As an illustrative example throughout the article, we will use the simple SIR epidemic model. At any time, each individual is either susceptible (S), infectious (I), or recovered (R). In this model, there are two mechanistic parameters of interest that govern the transitions of individuals between states S, I, and R : the transmission rate of the pathogen λ and the recovery rate γ. More precisely, individuals can move from state S to I according to λ, or from state I to R according to γ (Figure 2.1).

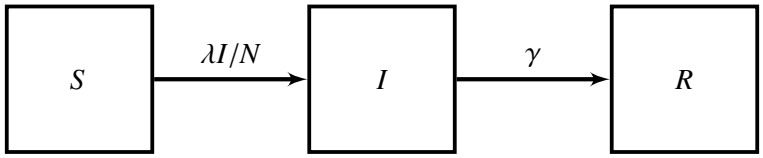

FIGURE 2.1 – SIR compartmental model with three blocks corresponding respectively to susceptible (S), infectious (I), and recovered (R) individuals. Transitions of individuals from state S to I are governed by the transmission rate λ, and transitions of individuals from state I to R are governed by the recovery rate γ of the epidemic.

One of the main goals of epidemic studies is to estimate such mechanistic parameters from the available data. One of the most natural probabilistic representations of compartmental epidemic models is the continuous-time Markov jump process (see Section 2.2.2). Inference for Markov jump processes is straightforward when sample paths are completely observed. In the context of epidemics, this is equivalent to the observation of all infection and recovery times for all individuals in the population. This rarely occurs in practice ; often one or more of the coordinates (i.e., $S(t)$, $I(t)$) are not observed, and available observations are only collected at discrete time points t_k with $0 = t_0 < t_1 < t_2 < \cdots < t_n = T$ over a finite time interval $[0, T]$. More specifically, the data often consists of counting newly infected individuals $N_I(t_k)$ on successive time intervals $[t_{k-1}, t_k]$. Alternatively, the successive numbers of infectious individuals $I(t_k)$ are sometimes available, especially for low population sizes. Moreover, it is common that the available data is affected by several sources of noise such as under-reporting of infection events or—when reported—imperfect diagnostic tests. Essentially, the nature of such data makes it difficult to infer key epidemic parameters : (i) observations are available at discrete time points, (ii) not all coordinates of the dynamical model are observed, and (iii) systematic reporting and measurement errors have to be taken into account.

2.2.2 Approximation of large population epidemic models and the autoregressive point of view

Consider an epidemic in a closed population with homogeneous mixing modeled by a d-dimensional Markov jump process $\mathcal{Z}(t)$, where d is the number of compartments corresponding to successive health statuses within the population. If N is the population size, the state space of $(\mathcal{Z}(t), t \geq 0)$ is $E = \{0, \ldots, N\}^d$. Let $Q = (q_{k,l}, k, l \in E)$ denote its Q-matrix ; the latter satisfies $\forall l \neq k,\ q_{k,l} \geq 0$, and $q_{k,k} = -\sum_{l \in E, l \neq k} q_{k,l}$. There are two standard ways of describing this jump process (see e.g., Norris (1997)) :

(i) By the underlying jump chain and holding times. Starting from Q, set $\pi_{k,l} = \frac{q_{k,l}}{q_k}$ with $q_k = -q_{k,k}$, $\pi_{k,k} = 0$ if $q_k \neq 0$, and $\pi_{k,k} = 1$ if $q_k = 0$. The process stays in state k according to an exponential distribution $\mathcal{E}(q_k)$ and jumps to state l with probability $\pi_{k,l}$.

(ii) Using its infinitesimal generator : as $h \to 0$, $\mathbb{P}(\mathcal{Z}(t+h) = l | \mathcal{Z}(t) = k) = \delta_{k,l} + q_{k,l}h + o_P(h)$, where $\delta_{k,l}$ denotes the Kronecker function ($\delta_{k,l} = 1$ if $l = k$, $\delta_{k,l} = 0$ if $l \neq k$).

Hence, for f a measurable function $E \to \mathbb{R}$, if $\mathbb{E}_k$ denotes the expectation conditional on $\mathcal{Z}(0) = k$, $[Qf](k) = \sum_{l \in E} q_{k,l} f(l) = \lim_{t \to 0} \frac{1}{t}(\mathbb{E}_k f(\mathcal{Z}(t)) - f(k))$. Simulations of $\mathcal{Z}(t)$ are usually based on (i), while (ii) relies on general properties of Markov processes.

For any vector V or matrix M, let V^t or M^t denote their transpose. For a jump $\ell \neq (0, \dots, 0)^t$ of $\mathcal{Z}(t)$, we define the jump function :

$$\alpha_\ell(k) = q_{k,k+\ell} \quad \text{for} \quad k, k+\ell \in E.$$

Consider now the normalized Markov jump process $(\mathcal{Z}_N(t))_{t \geq 0}$:

$$\mathcal{Z}_N(t) = \frac{\mathcal{Z}(t)}{N} \in E^N = \{k/N, k \in E\}. \tag{2.1}$$

The associated jump functions are, for $x \in E^N$, $\alpha_\ell^N(x) = \frac{1}{N}\alpha_\ell([Nx])$. Assume that the process $(\mathcal{Z}(t))$ is density-dependent, i.e.,

$$\begin{aligned} &\textbf{H1 :} \ \forall \ell, \quad \forall x \in [0,1]^d, \ \frac{1}{N}\alpha_\ell([Nx]) \underset{N \to +\infty}{\to} \beta_\ell(x), \\ &\textbf{H2 :} \ \forall \ell, \quad \beta_\ell \in C^2([0,1]^d, \mathbb{R}), \end{aligned}$$

where $[Nx]$ is the vector of integers $[Nx_1], \dots, [Nx_d]$, with $[Nx_i]$ the integer part of Nx_i. Next, define for $x \in [0,1]^d$ the function $b(\cdot)$ and the $d \times d$ symmetric non-negative matrix $\Sigma(\cdot)$:

$$b(x) = \sum_{\ell \in E^-} \ell \, \beta_\ell(x) \ ; \quad \Sigma(x) = \sum_{\ell \in E^-} \beta_\ell(x) \, \ell \, \ell^t. \tag{2.2}$$

For the SIR epidemic model in a closed population, we have that $S(t) + I(t) + R(t) = N$ for all t. Therefore, its state space is $E = \{0, \dots, N\}^2$. Only two jumps are possible from $k = (S, I)^t$:

— $\ell_1 = (-1, +1)^t : (S, I) \to (S-1, I+1) \Rightarrow q_{k,k+\ell_1} = \lambda S I / N = \alpha_{\ell_1}(k)$,

— $\ell_2 = (0, -1)^t : (S, I) \to (S, I-1) \Rightarrow q_{k,k+\ell_2} = \gamma I = \alpha_{\ell_2}(k)$.

This process is density dependent : if $s = \frac{S}{N}, i = \frac{I}{N}$, then $\frac{1}{N}\alpha_{\ell_1}([Ns], [Ni]) = \frac{1}{N}(\lambda [Ns]) \frac{[Ni]}{N} \to \lambda s i$ and $\frac{1}{N}\alpha_{\ell_2}([Ns], [Ni]) = \frac{1}{N}\gamma [Ni] \to \gamma i$ as $N \to \infty$.

Moreover (2.2) is, for $x = \begin{pmatrix} s \\ i \end{pmatrix}$, $b(x) = \lambda s i \begin{pmatrix} -1 \\ 1 \end{pmatrix} + \gamma i \begin{pmatrix} 0 \\ -1 \end{pmatrix}$, $\Sigma(x) = \lambda s i \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \gamma i \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$.

We now recall the law of large numbers result stated (for instance) in Britton and Pardoux (2020).

Lemma 1. *Assume that $(\mathcal{Z}(t))$ satisfies (H1), (H2), and $\mathcal{Z}_N(0) \to x_0$ as $N \to +\infty$. Then, $(\mathcal{Z}_N(t))$ converges almost surely uniformly on $[0,T]$ to the solution $x(t)$ of the ordinary differential equation*

$$\frac{dx}{dt} = b(x(t)); \quad x(0) = x_0. \tag{2.3}$$

If $x_0 = (0, \dots, 0)^t$, then $x(t) = 0$ for all t and (2.3) no longer adequately describes the epidemic dynamics (see e.g., Britton and Pardoux (2020) Part I). Equation (2.3) describes the dynamics in the case of a major outbreak corresponding to $x_0 \neq (0, \dots, 0)^t$.

In Guy et al. (2015), by extending the results of Ethier and Kurtz (2005), another approximation of the epidemic model was proposed, leading to a diffusion process $(Z_N(t))_{t\geq 0}$ with the small diffusion matrix $\frac{1}{N}\Sigma(x)$, where Σ is the matrix defined in (2.2) :

$$\begin{cases} dZ_N(t) &= b(Z_N(t)) + \frac{1}{\sqrt{N}}\sigma(Z_N(t))\, dB(t), \\ Z_N(0) &= x_0, \end{cases} \tag{2.4}$$

where $(B(t))_{t\geq 0}$ is a d-dimensional Brownian motion and σ a $d \times d$ matrix such that

$$\sigma(x)\sigma^t(x) = \Sigma(x). \tag{2.5}$$

For stochastic differential equations with small noise (i.e., proportional to $\frac{1}{N}$), an approximation of $Z_N(t)$ can be obtained using (2.2)-(2.5), based on the theory of perturbations of dynamical systems (see e.g., Azencott (1982), Freidlin and Wentzell (1978)) :

$$\begin{cases} Z_N(t) &= x(t) + \frac{1}{\sqrt{N}}g(t) + \frac{1}{\sqrt{N}}R_N(t), \\ dg(t) &= \nabla_x b(x(t))\; g(t)\; dt + \sigma(x(t))\; dB(t)\,; \quad g(0) = 0, \\ \text{with} & \sup_t \|R_N(t)\| \to 0 \;\text{ in probability as } N \to +\infty, \end{cases} \tag{2.6}$$

where $\nabla_x b(x)$ denotes the matrix $(\frac{\partial b_i}{\partial x_j}(x))_{1\leq i,j\leq d}$. The stochastic differential equation for $g(\cdot)$ defined in (2.6) can be solved explicitly (see e.g., Guy et al. (2014) for details) and its solution is the time-inhomogeneous Gaussian process

$$g(t) = \int_0^t \Phi(t,s)\sigma(x(s))\; dB(s), \tag{2.7}$$

where $\Phi(t,s)$ satisfies $\dfrac{\partial \Phi}{\partial t}(t,s) = \nabla_x b(x(t))\Phi(t,s)$, $\Phi(s,s) = I_d$. Hence, $\Phi(t,s)$ is the $d \times d$ matrix

$$\Phi(t,s) = \exp\left(\int_s^t \nabla_x b(x(u))\; du\right). \tag{2.8}$$

Using (2.3) and (2.7), let us define the Gaussian process $G_N(t)$:

$$G_N(t) = x(t) + \frac{1}{\sqrt{N}}g(t). \tag{2.9}$$

Consider now the Wasserstein-1 distance on the interval $[0,T]$ between $\mathbb{R}^d$-valued processes U_t, V_t on $[0,T]$. $W_{1,T}(U,V) = \inf \mathbb{E}(\|U-V\|_T)$, where if $x : [0,T] \to \mathbb{R}^d$, $\|x\|_T = \sup_{0\leq t\leq T} \|x(t)\|$, and the above infimum is over all couplings of two processes. According to Britton and Pardoux (2020), Part I, Theorem 2.4.1, the following holds.

Proposition 1. *For all $T > 0$, the Wasserstein-1 distances on $[0,T]$ between the three processes $(\mathcal{Z}_N(\cdot))$, $(Z_N(\cdot))$, and $(G_N(\cdot))$ defined in (2.1),(2.4),(2.9) satisfy, as $N \to \infty$,*

$$\sqrt{N}W_{1,T}(\mathcal{Z}_N, Z_N) \to 0, \quad \sqrt{N}W_{1,T}(\mathcal{Z}_N, G_N) \to 0, \quad \text{and } \sqrt{N}W_{1,T}(Z_N, G_N) \to 0.$$

From a statistical point of view, this proposition has important consequences : given the fact that these distances are $o(N^{-1/2})$, we develop our inference method by plugging the observations into the likelihood of either the diffusion process (Z_N) or the Gaussian process (G_N). This approach is often used to derive approximate likelihoods or contrasts for stochastic processes. For instance, for discretely observed diffusion processes, parametric inference is often based on the likelihood of the Euler scheme of the diffusion (see e.g., Kessler et al. (2012)). Moreover, it was proved in

Guy et al. (2014) that parametric inference based on (G_N) leads to efficient estimators for the parameters ruling the jump process.

From here on, we will use the approximation of $(\mathcal{Z}_N)$ by the Gaussian process (G_N). Let us now consider a parametric model for epidemic dynamics. This yields a parametric continuous-time approximate model for epidemic dynamics, with parameter

$$\eta = (\zeta, x_0), \tag{2.10}$$

where ζ contains the parameters found in the transition rates of the jump process, and therefore in the functions $\beta_\ell(x)$ defined in **H1**, and x_0 is the initial point of the ordinary differential equation (ODE) defined in Lemma 1. As mentioned in Section 2.2.1, the process is however observed at discrete times t_k, where (t_k) is an increasing sequence on $[0, T]$, with $t_0 = 0 < t_1 \cdots < t_n = T$. We therefore deduce from above a discrete-time representation of the epidemic evolution.

Let us denote by $\mathcal{F}_t = \sigma(B(s), s \leq t)$. Then the following holds.

Proposition 2. *There exists a sequence of independent Gaussian random variables (U_k) such that*

(i) For all k, U_k is $\mathcal{F}_{t_k}$-measurable and independent of $\mathcal{F}_{t_{k-1}}$.

(ii) The process G_N defined in (2.9) is an AR(1) process and satisfies, using (2.3), (2.8), $G_N(0) = x_0$, for $k \geq 1$,

$$G_N(t_k) = F_k(\eta) + A_{k-1}(\eta)\, G_N(t_{k-1}) + U_k,$$

where

$$\begin{aligned} A_{k-1}(\eta) &= A(\eta, t_{k-1}) = \Phi(\eta, t_k, t_{k-1}), \\ F_k(\eta) &= F(\eta, t_k) = x(\eta, t_k) - \Phi(\eta, t_k, t_{k-1}) x(\eta, t_{k-1}), \end{aligned}$$

and (U_k) are independent random variables such that

$$U_k \sim \mathcal{N}_d(0, T_k(\eta)),$$

with

$$T_k(\eta) = \frac{1}{N} \int_{t_{k-1}}^{t_k} \Phi(\eta, t_k, s) \Sigma(\eta, x(\eta, s))\, \Phi^t(\eta, t_k, s) ds.$$

The proof of Proposition 2 is given in the Appendix. Using now that $\sup_t \|\mathcal{Z}_N(t) - G_N(t)\| = \frac{1}{\sqrt{N}} o_P(1)$, Proposition 2 becomes, setting $X_k := X(t_k) = \mathcal{Z}_N(t_k)$, $X_0 = x_0$, for $k \geq 1$,

$$X_k = F_k(\eta) + A_{k-1}(\eta) X_{k-1} + U_k. \tag{2.11}$$

2.2.3 Approximation of the observation model

Assume now that there are noisy observations $O(t_k)$ of the original jump process $\mathcal{Z}(t)$ (with state space $E = \{0, \ldots, N\}^d$ at discrete times t_k). As mentioned in Section 2.2.1, it often occurs in practice that not all epidemiological health states are observed. We account for this by introducing a projection operator $B : \mathbb{R}^d \to \mathbb{R}^q$ with $q \leq d$, where $BX(\cdot)$ contains only the coordinates that can be observed. Therefore B is a $d \times q$ matrix whose elements are 0 and 1. For $k = 0, \ldots, n$, define

$$C(t_k) = (C_1(t_k), \ldots, C_q(t_k))^t = B\mathcal{Z}(t_k) \in \{0, \ldots, N\}^q.$$

In an initial approach, assume that each component of C is observed with independent reporting rate p_i and measurement errors. In this way, we propose a rather general model for the observations conditional on $\mathcal{Z}(t)$, for $1 \leq i \leq q$:

$$O_i(t_k) = O_{i,1}(t_k) + O_{i,2}(t_k), \text{ with } O_{i,1}(t_k) \sim \text{Binomial}(C_i(t_k), p_i), \; O_{i,2}(t_k) \sim \mathcal{N}\left(0, \tau_i^2 C_i(t_k)\right), \quad (2.12)$$

where, conditional on $\sigma(\mathcal{Z}(s), 0 \leq s \leq t_k)$, the variables $O_{i,1}(t_k)$ and $O_{i,2}(t_k)$ are independent. This yields a new higher-dimensional parameter containing parameters for both the epidemic (i.e., η defined in (2.10)) and observation processes :

$$\theta = (\eta, (p_1, \ldots, p_q), (\tau_1^2, \ldots, \tau_q^2)).$$

Consider now the normalized process $\mathcal{Z}_N(t)$. We can then define $C_N(t) = B\mathcal{Z}_N(t)$ and associated normalized observations $O_N(t_k) = \frac{1}{N} O(t_k)$. A Gaussian approximation of the observation process has first and second moments which satisfy

$$\begin{aligned} E(O_{N,i}(t_k)|\mathcal{Z}(t_k)) &= p_i C_{N,i}(t_k), \\ Var(O_{N,i}(t_k)|\mathcal{Z}(t_k)) &= \frac{1}{N}(p_i(1-p_i) + \tau_i^2) C_{N,i}(t_k). \end{aligned}$$

Using now (2.6) and Proposition 1, we get that

$$C_N(t) = B\mathcal{Z}_N(t) = Bx(\eta, t) + \frac{1}{\sqrt{N}} Bg(\eta, t) + \frac{1}{\sqrt{N}} o_P(1).$$

The Gaussian process $g(\eta, t)$ is uniformly bounded in probability on $[0, T]$, so we have that

$$Var(O_{N,i}(t_k)|\mathcal{Z}(t_k)) = \frac{1}{N}(p_i(1-p_i) + \tau_i^2)(Bx(\eta, t_k))_i + O_P(N^{-3/2}).$$

Let us next define the q-dimensional matrices

$$P(\theta) = \text{diag}(p_i)_{1 \leq i \leq q}, \quad Q_k(\theta) = \frac{1}{N} \text{diag}\left((p_i(1-p_i) + \tau_i^2)(Bx(\eta, t_k))_i\right), \quad (2.13)$$

and the $q \times d$ matrix

$$B(\theta) = P(\theta)B.$$

The Gaussian approximations (Y_k) of the observations $O_N(t_k)$ satisfy that conditionally on $\mathcal{Z}(t_k)$,

$$Y_k = B(\theta)X_k + V_k \;\text{ with }\; V_k \sim \mathcal{N}_q(0, Q_k(\theta)), \quad (2.14)$$

where (V_k) are independent random variables such that for all k, V_k is independent of $\mathcal{Z}_N(t_k)$.

2.2.4 Application on the SIR epidemic model

Let us now illustrate the model approximations derived in Sections 2.2.2 and 2.2.3 on the simple SIR model introduced in Section 2.2.1. The Markov jump process $\mathcal{Z}(t) = (S(t), I(t))^t$, $t \geq 0$ is defined in Section 2.2.2. The parameters controlling the dynamics of the system are

$$\eta = (\lambda, \gamma, x_0) = (\lambda, \gamma, s_0, i_0),$$

which include the transition rates λ and γ, and the initial point $x_0 = (s_0, i_0)$ (cf Lemma 1).

Dynamical state model Let us define the key quantities necessary to derive the appropriate Gaussian process $(G_N(t))$ as defined in (2.9), including the dependence on η :

$$G_N(t) = x(\eta, t) + \frac{1}{\sqrt{N}} g(\eta, t).$$

The first important element is $x(\eta, t) = (s(\eta, t), i(\eta, t))^t$, solution of the following ODEs :

$$\begin{cases} \frac{ds}{dt}(\eta, t) & = -\lambda s(\eta, t) i(\eta, t), \\ \frac{di}{dt}(\eta, t) & = \lambda s(\eta, t) i(\eta, t) - \gamma i(\eta, t), \\ x_0 & = (s_0, i_0). \end{cases}$$

When there is no ambiguity, we denote by s and i respectively $s(\eta, t)$ and $i(\eta, t)$. Then, to get $g(\eta, \cdot)$, we need to derive the functions $b(\eta, \cdot)$ and $\Sigma(\eta, \cdot)$ from (2.2) (see Section 2.2.2) :

$$b(\eta, s, i) = \begin{pmatrix} -\lambda si \\ \lambda si - \gamma i \end{pmatrix}; \quad \Sigma(\eta, s, i) = \begin{pmatrix} \lambda si & -\lambda si \\ -\lambda si & \lambda si + \gamma i \end{pmatrix}, \tag{2.15}$$

and the Cholesky decomposition of $\Sigma(\eta, \cdot)$:

$$\sigma(\eta, s, i) = \begin{pmatrix} \sqrt{\lambda si} & 0 \\ -\sqrt{\lambda si} & \sqrt{\gamma i} \end{pmatrix}.$$

From (2.15), we deduce the gradient of b :

$$\nabla_x b(\eta, s, i) = \begin{pmatrix} -\lambda i & -\lambda s \\ \lambda i & \lambda s - \gamma \end{pmatrix},$$

and the resolvent matrix defined in (2.8) :

$$\Phi(\eta, t, s) = \exp\left(\int_s^t \nabla_x b(\eta, x(\eta, u)) du \right).$$

Finally, we obtain

$$g(\eta, t) = \int_0^t \Phi(\eta, t, u) \sigma(\eta, x(\eta, u)) dB(u),$$

where $(B(u))_{u \geq 0}$ is a bidimensionnal Brownian motion.

Discrete-time system For simplicity, we assume a regular sampling : $t_k = k\Delta$, $k = 0, \ldots, n$, $T = n\Delta$. The dependence with respect to Δ is explicitly given in the equations. The approximate autoregressive model, setting $X_k = \mathcal{Z}_N(t_k) = (S_N(k\Delta), I_N(k\Delta))^t$, is given by :

$$\begin{cases} X_k & = F_k(\eta, \Delta) + A_{k-1}(\eta, \Delta) X_{k-1} + U_k, \quad \text{where} \\ F_k(\eta, \Delta) & = x(\eta, t_k) - \Phi(\eta, t_k, t_{k-1}) x(\eta, t_{k-1}), \quad A_{k-1}(\eta, \Delta) = \Phi(\eta, t_k, t_{k-1}), \\ U_k \sim \mathcal{N}_2(0, T_k(\eta, \Delta)) & \quad \text{with } T_k(\eta, \Delta) = \frac{1}{N} \int_{t_{k-1}}^{t_k} \Phi(\eta, t_k, s) \Sigma(\eta, x(\eta, s)) \Phi^t(\eta, t_k, s) ds. \end{cases} \tag{2.16}$$

Observation model Suppose for example that only the infected individuals are observed with reporting and measurement errors. This corresponds to considering in (2.12) :

$$O_1(t_k) \sim \text{Binomial}(I(t_k), p), \quad O_2(t_k) \sim \mathcal{N}(0, \tau^2 I(t_k)). \tag{2.17}$$

Hence the full parameter vector is $\theta = (\lambda, \gamma, s_0, i_0, p, \tau^2)$. To derive (2.14) from this example, we define the operator $B(\theta) = pB$, where $B : (x_1, x_2)^t \to x_2$ is the projection operator on the infected compartment, and $Q_k(\theta) = \frac{1}{N}(p(1-p) + \tau^2) i(\eta, t_k)$, with Q_k is defined in (2.13).

By joining (2.16) with the Gaussian approximate observation model defined above, we get the following discrete-time state-space model :

$$\begin{cases} X_k = & F_k(\eta,\Delta) + A_{k-1}(\eta,\Delta)X_{k-1} + U_k, \quad \text{with } U_k \sim \mathcal{N}_2\left(0, T_k(\eta,\Delta)\right), \\ Y_k = & p\begin{pmatrix} 0 & 1 \end{pmatrix} X_k + V_k, \quad \text{with } V_k \sim \mathcal{N}\left(0, \frac{1}{N}(p(1-p)+\tau^2)i(\eta,t_k)\right). \end{cases}$$

2.3 Parameter estimation using Kalman filtering techniques

2.3.1 Approximate likelihood inference

The parameters of interest in the general case are denoted by $\theta = (\eta, (p_1,\dots,p_q), (\tau_1^2,\dots,\tau_q^2))$, where η contains the parameters controlling the dynamics and x_0, whereas $(p_1,\dots,p_q)$ and $(\tau_1^2,\dots,\tau_q^2)$ are derived from the reporting and measurements errors in the observations. Our aim is to estimate the unknown parameters θ from observations $y_{n:0} = (y_0,\dots,y_n)$ obtained at discrete time points $t_0 < t_1 < \cdots < t_n$. Joining (2.11) and (2.14), we get the following discrete-time Gaussian state-space setting that is more convenient for inference :

$$\begin{cases} X_k &= F_k(\eta) + A_{k-1}(\eta)X_{k-1} + U_k, \\ Y_k &= B(\theta)X_k + V_k, \end{cases} \tag{2.18}$$

where all quantities are explicitly defined in Sections 2.2.2 and 2.2.3. Using (2.18), we propose to estimate θ by maximizing the associated likelihood $L(\cdot; Y_0,\dots,Y_n)$:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ L(\theta; Y_0,\dots,Y_n). \tag{2.19}$$

The log-likelihood of the observations $y_0,\dots,y_n$ is given by :

$$\mathcal{L}(\theta; y_0,\dots,y_n) = \log f(\theta, y_0) + \sum_{k=1}^{n} \log f_k(\theta; y_k|y_{k-1:0}). \tag{2.20}$$

Computing $\mathcal{L}(\theta; y_0,\dots,y_n)$ requires the computation of the two first moments of the Gaussian conditional distributions corresponding to each $\log f_k(\theta;\dots)$ term. This relies on the computation of the predictive distributions $\nu_{k|k-1:0}(\theta; dx) = \mathcal{L}(X_k|y_{k-1:0})$, $k \geq 1$, from which we derive the conditional densities

$$f_k(\theta; y_k|y_{k-1:0}) = \int f(y_k|x)\nu_{k|k-1:0}(\theta; dx).$$

Usually, these conditional distributions are obtained by means of filtering methods, based on the iterative computations of the conditional distributions :

- the *predictive distribution* : $\mathcal{L}(X_k|y_{k-1},\dots,y_0) = \nu_{k|k-1:0}(dx)$, $k \geq 1$, with the convention $\nu_{0,0}(dx) = \mathcal{L}(X_0)$,
- the *updating distribution* : $\mathcal{L}(X_k|y_k,\dots,y_0) = \nu_{k|k:0}(dx)$, $k \geq 0$,
- the *marginal distribution* : $\mathcal{L}(Y_k|y_{k-1},\dots,y_0) = \mu_{k|k-1:0}$, $k \geq 1$, with the convention $\mu_{0|0:0}(dx) = \mathcal{L}(Y_0)$.

In the special case of the Gaussian state space model and Gaussian noise, all of these distributions are Gaussian and therefore characterized by their mean and covariance matrix. Using notation specific to Kalman filtering, let us set

$$\begin{aligned} \mathcal{L}(X_k|y_{k-1},\cdots,y_0) &= \nu_{k|k-1:0}(dx) = \mathcal{N}_d(\hat{X}_k, \hat{\Xi}_k) \quad (\textit{predictive distribution}). \\ \mathcal{L}(X_k|y_k,\cdots,y_0) &= \nu_{k|k:0}(dx) = \mathcal{N}_d(\bar{X}_k, \bar{T}_k) \quad (\textit{updating distribution}). \\ \mathcal{L}(Y_k|y_{k-1},\cdots,y_0) &= \mu_{k|k-1:0} = \mathcal{N}_q(\hat{M}_k, \hat{\Omega}_k) \quad (\textit{marginal distribution}). \end{aligned}$$

The Gaussian approximations defined in (2.11), (2.14) and (2.18) allow us to use specific properties of Gaussian distributions that are recalled below.

2.3.1.1 Preliminary results in the general framework of Kalman filtering

Let $(X_i, i \geq 0)$ be a non-centered d-dimensional Gaussian $AR(1)$ process and assume that only q coordinates of (X_i) are observed, with Gaussian noise. Computations of the conditional distributions rely on a Kalman filter approach, which is derived from the following lemma.

Lemma 2. *Assume that X is a random variable with distribution $\mathcal{N}_d(\xi, T)$ which conditional on X, Y has distribution $\mathcal{N}_q(BX, Q)$. Then, $\mathcal{L}(X|Y)$ is Gaussian : $\mathcal{N}_d(\bar{\xi}(y), \bar{T})$, with*

$$\bar{\xi}(y) = \xi + TB^t(BTB^t + Q)^{-1}(y - B\xi); \quad \bar{T} = T - TB^t(BTB^t + Q)^{-1}BT. \tag{2.21}$$

Remark 1. *We stress that Lemma 2 holds even if Q is singular. In particular, the formula holds when $Q = 0$ and B is a projection operator, i.e., the observations are $Y_k = BX_k$, provided that T is non-singular.*

Let us now go back to our general setting (X_k, Y_k) defined in (2.18).

Proposition 3. *Assume that (X_k, Y_k) are defined as in (2.18). Then, $\nu_{k|k-1:0}(dx)$, $\nu_{k|k:0}(dx)$, and $\mu_{k|k-1:0})(dy)$ satisfy, with the initialization $\hat{X}_0 = \xi_0$, $\hat{\Xi}_0 = T_0$, for $k \geq 0$,*

(i) Prediction : $\nu_{k|k-1:0}(dx) \sim \mathcal{N}_d(\hat{X}_k, \hat{\Xi}_k)$ with
$\hat{X}_k = F_k + A_{k-1}\bar{X}_{k-1}$, $\hat{\Xi}_k = A_{k-1}\bar{T}_{k-1}A_{k-1}^t + T_k$.

(ii) Updating : $\nu_{k|k:0}(dx) \sim \mathcal{N}_d(\bar{X}_k, \bar{T}_k)$ with
$\bar{X}_k = \hat{X}_k + \hat{\Xi}_k B^t(B\hat{\Xi}_k B^t + Q_k)^{-1}(Y_k - B\hat{X}_k)$, $\bar{T}_k = \hat{\Xi}_k - \hat{\Xi}_k B^t(B\hat{\Xi}_k B^t + Q_k)^{-1}B\hat{\Xi}_k$.

(iii) Marginal distribution : $\mu_{k+1|k:0}(dy) \sim \mathcal{N}_q(\hat{M}_{k+1}, \hat{\Omega}_{k+1})$ with
$\hat{M}_{k+1} = B\hat{X}_{k+1}$, $\hat{\Omega}_{k+1} = B\hat{\Xi}_{k+1}B^t + Q_{k+1}$.

Using specific notation from Kalman filtering, we recover a modified version of the Kalman algorithm. Assume that $X_0 \sim \mathcal{N}_d(\xi_0, T_0)$ and that, for all $k \geq 0$, the matrices Γ_k defined below are non-singular. Then, setting $\hat{X}_0 = \xi_0, \hat{\Xi}_0 = T_0$, we have

$$\begin{array}{rcll}
\epsilon_{k-1} & = & Y_{k-1} - B\hat{X}_{k-1}, & \textit{(innovation)} \\
\Gamma_{k-1} & = & B\hat{\Xi}_{k-1}B^t + Q_{k-1}, & \textit{(innovation covariance)} \\
H_{k-1} & = & A_{k-1}\hat{\Xi}_{k-1}B^t\Gamma_{k-1}^{-1}, & \textit{(Kalman gain)} \\
\hat{X}_k & = & F_k + A_{k-1}\hat{X}_{k-1} + H_{k-1}\epsilon_{k-1}, & \textit{(predicted mean state estimation)} \\
\hat{\Xi}_k & = & (A_{k-1} - H_{k-1}B)\hat{\Xi}_{k-1}A_{k-1}^t + T_k. & \textit{(predicted error covariance)}
\end{array}$$

Therefore, the marginal distributions appearing in the computation of the log-likelihood (2.20) are $\mu_{k+1|k:0}(dy) \sim \mathcal{N}_q(\hat{M}_{k+1}, \hat{\Omega}_{k+1})$, with

$$\hat{M}_{k+1} = B\hat{X}_{k+1}, \quad \hat{\Omega}_{k+1} = B\hat{\Xi}_{k+1}B^t + Q_{k+1}. \tag{2.22}$$

2.3.1.2 Recursive computation of the approximate log-likelihood

An important consequence of the previous section is that we can compute (2.20) based on the recursive computations of the first moments of the Gaussian distributions corresponding to each term of the log-likelihood. By explicitly accounting for the dependence on θ of moments given in (2.22), we obtain :

$$\mathcal{L}(\theta; y_0, \ldots, y_n) = C + \log f(\theta; y_0) - \frac{1}{2}\sum_{k=1}^{n}\left[\log\left(|\hat{\Omega}_k(\theta)|\right) + (y_i - \hat{M}_k(\theta))^t\left(\hat{\Omega}_k(\theta)\right)^{-1}(y_i - \hat{M}_k(\theta))\right],$$

with C a constant (independent of the parameters) and $|A|$ denoting the determinant of the matrix A.

Note that the sampling interval Δ plays an important role in the various key quantities involved in the Kalman recursions (see A.1 for details).

2.3.2 Application on the SIR epidemic model

Let us again take the example of SIR epidemics, when only the infected individuals are observed with reporting and measurement errors, considered in Section 2.2.4. By assuming an initial distribution $X_0 \sim \mathcal{N}_2(\xi_0, T_0)$, setting $\hat{X}_0 = \xi_0, \hat{\Xi}_0 = T_0$, and applying the algorithm given in Proposition 3, we have, for $k = 0, \ldots, n-1$:

$$\begin{aligned}
\epsilon_{k-1}(\theta) &= Y_{k-1} - p\hat{I}_{k-1}(\theta), && \text{(scalar)}\\
\Gamma_{k-1}(\theta) &= p^2(\hat{\Xi}_{k-1}(\theta))_{22} + \tfrac{1}{N}(p(1-p)+\tau^2)i(\eta, t_{k-1}), && \text{(scalar)}\\
H_{k-1}(\theta) &= pA_{k-1}(\eta)\hat{\Xi}_{k-1}(\theta)\begin{pmatrix}0\\1\end{pmatrix}\Gamma_{k-1}^{-1}(\theta), && \text{(vector)}\\
\hat{X}_k(\theta) &= F_k(\eta) + A_{k-1}(\eta)\hat{X}_{k-1}(\theta) + H_{k-1}(\theta)\epsilon_{k-1}(\theta), && \text{(vector)}\\
\hat{\Xi}_k(\theta) &= \left(A_{k-1}(\eta) - pH_{k-1}(\theta)\begin{pmatrix}0 & 1\end{pmatrix}\right)\hat{\Xi}_{k-1}(\theta)A_{k-1}(\eta)^t + T_k(\eta). && (2\times 2 \text{ matrix})
\end{aligned}$$

This yields the marginal distributions :

$$\hat{M}_{k+1}(\theta) = p\hat{I}_{k+1}(\theta), \quad \hat{\Omega}_{k+1}(\theta) = p^2\left(\hat{\Xi}_{k+1}(\theta)\right)_{22} + \frac{1}{N}(p(1-p)+\tau^2)i(\eta, t_{k+1})),$$

which are used to compute the likelihood

$$\mathcal{L}(\theta, y_1, \ldots, y_n) \simeq -\frac{1}{2}\sum_{k=1}^{n}\log\hat{\Omega}_k(\theta) - \frac{1}{2}\sum_{k=1}^{n}\frac{(y_k - \hat{M}_k(\theta))^2}{\hat{\Omega}_k(\theta)}.$$

2.4 Simulation study

We assessed the performance of our method on simulated SIR epidemics in which only the infectious compartment is observed at discrete time points (see Section 2.2.4 where the model is fully described).

2.4.1 Simulation settings

Data simulation We first simulated SIR dynamics according to the Markov jump process using the Gillespie algorithm (Gillespie (1977)). Only trajectories that did not exhibit early extinction were considered for inference. The theoretical proportion of these trajectories is given by $1 - (\gamma/\lambda)^{I_0}$ (Andersson and Britton (2000)), where I_0 is the number of infectious individuals at time 0. We simulated two cases. First, for the emergent trajectories, the observations were generated by binomial draws from $I(t)$ at $n+1$ discrete time points $t_0 < t_1 < \ldots < t_n$. In (2.17), this amounts to considering $\tau = 0$, with simulated observations finally obtained via $O(t_k) = O_1(t_k) \sim \text{Binomial}(I(t_k), p)$. Second, we considered the more general case where observations are $O(t_k) = O_1(t_k) + O_2(t_k)$, with $O_1(t_k) \sim \text{Binomial}(I(t_k), p)$, $O_2(t_k) \sim \mathcal{N}\left(0, \tau^2 I(t_k)\right)$, where the non-zero measurement error τ is an additional parameter to estimate. Figure 2.2 represents epidemic trajectories corresponding to the various steps of data simulation. These plots illustrate the variability in the stochastic trajectories compared to the deterministic counterpart of the SIR model, and the loss of information from the unobservable real dynamics to the observations available for inference. Moreover, the second source of error, driven by the measurement error τ, seems to have a minor impact on the global observational noise compared to the reporting error. The evolution of the number of susceptible individuals is not shown in Figure 2.2. From the point of view of inference, the S compartment is a latent variable, the observations being only available for the infected state.

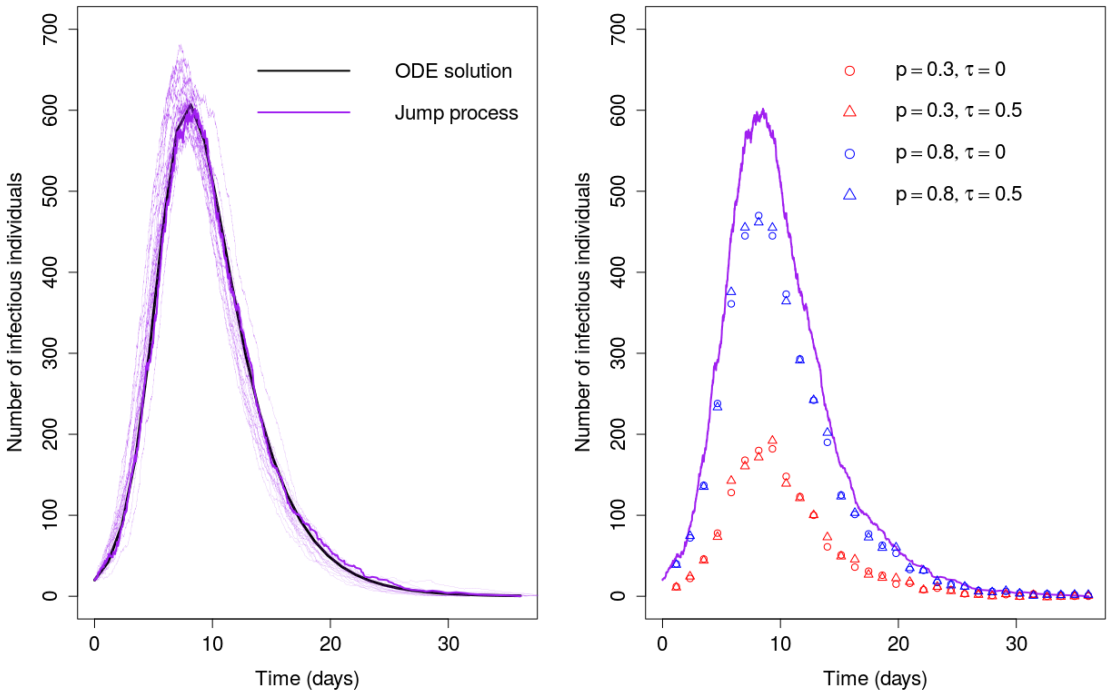

FIGURE 2.2 – Left panel : ODE solution for the number of infected individuals I (plain black line) and 20 trajectories of the Markov jump process for I (purple lines) when $N = 2000$. Right panel : $n = 30$ observations obtained from a particular trajectory of the jump process (in bold purple in the left panel) as a function of time. The points and triangles stand for observations generated with measurement error terms $\tau = 0$ and $\tau = 0.5$ respectively, and the blue and red symbols represent observations generated with $p = 0.8$ and $p = 0.3$ respectively.

Numerical scenarios We used the following parameter values for the simulation of the epidemics : $\lambda = 1, \gamma = 1/3$, and initial starting points $S(0)/N = s_0 = 0.99$, $I(0)/N = i_0 = 0.01$, $R(0) = 0$ (hence with $s_0 + i_0 = 1$). Observations were generated under two scenarios : *i)* high reporting rate $p = 0.8$ and *ii)* low reporting rate $p = 0.3$. Two experiments were considered concerning the measurement error : $\tau = 0$ (experiment 1) and $\tau = 0.5$ (experiment 2). Scenarios combining three population sizes ($N \in \{1000, 2000, 10000\}$) with different values for the number of observations (n) for each epidemic trajectory were also investigated. For each value of N, conditionally on non extinction, 500 SIR epidemic dynamics were simulated. Observations were generated at regularly-spaced time points $t_k = k\Delta$ using, for a given scenario, the same value of Δ for each of the 500 epidemics (obtained by dividing the mean epidemic duration over 500 trajectories by a target number of observations n). As the epidemic duration is stochastic, we considered slightly different observation intervals $[0, T]$ for each epidemic and set the value of T as the first time point when the number of infected individuals became zero. This generates slightly different numbers of observations per epidemic trajectory.

2.4.2 Inference : settings, performance comparison, and implementation

The unknown parameters to be estimated are either $\theta = (\lambda, \gamma, p, i_0)$ or $\theta = (\lambda, \gamma, p, i_0, \tau)$, according to the experiment. Here, we do not need to estimate s_0 as $s_0 = 1 - i_0$. When $\tau \neq 0$, the

observational model used for the two estimation methods was a Gaussian model given as the sum of the two sources of noise in the data (reporting : Gaussian approximation of a binomial model ; measurement : Gaussian model). For each simulated dataset, θ is estimated with our Kalman filter-based estimation method (KM) and with the MIF algorithm (Ionides et al. (2006), Ionides et al. (2011), Ionides et al. (2015)), which is widely used in practice for statistical inference of epidemics. The simulation study was performed with the `R` software on a Bi-pro Xeon E5-2680 processor with 2.8 Ghz, 96 Go RAM, and 20 cores. MIF estimation was performed with the `mif2` function of the `POMP`-package (King et al. (2017)). We provide user-friendly code on the RunMyCode website (see A.6 for details).

Let us make some initial remarks on the algorithms and their practical implementations. Regardless of the estimation method used, maximisation of the log-likelihood requires considering several constraints : (i) strict positivity of λ, γ, i_0, (ii) $s_0+i_0 = 1$ (or $s_0+i_0 \leq 1$ in the general case), and (iii) $0 < p \leq 1$. To facilitate optimization, a different parameterization was implemented : $\lambda = \exp(\mu_1)$, $\gamma = \exp(\mu_2)$, $p = (1+\exp(\mu_3))^{-1}$, $i_0 = (1+\exp(\mu_4))^{-1}$, where $\mu_1, \mu_2, \mu_3, \mu_4 \in \mathbb{R}$. With no constraints on this new set of parameters, numerical optimization was more stable in practice.

The approximated log-likelihood given by Kalman filtering techniques cannot be maximized explicitly. We instead used the Nelder-Mead method implemented in the `optim` function in `R`, which requires inputting initial values for the unknown parameters. According to the amount of information available in the observations, the result of the optimization is more or less sensitive to these initial points. The same problem can occur for the MIF algorithm. The dependence on the initialization can be circumvented by trying different starting values (10 in the present case) and choosing the maximum value for the log-likelihood among them. The starting parameter values for the maximization algorithm were uniformly drawn from a hypercube encompassing the likely true values.

When the time intervals Δ between observations are large (which often occurs for low values of n), we computed the resolvent matrix defined in (2.8) as in (6.1) in order to obtain the approximated log-likelihood with Kalman filtering techniques.

MIF, based on particle filtering, returns an estimate of the log-likelihood of the observations by using resampling techniques. The parameter space is investigated by randomly perturbing the parameters of interest at each iteration, the amplitude of the perturbation decreasing as the iterations progress. The MIF algorithm has a complexity of $O(JM)$, where J and M are respectively the number of particles and the number of iterations. Running MIF requires specifying several tuning parameters. For the present study, the best results were obtained using $M = 100$ iterations, $J = 500$ particles, standard deviation `rw.sd` equal to 0.2 for the random walk for each parameter, and a cooling of the perturbations of `cooling.fraction.50=0.05` in the `POMP`-package (we drew inspiration from Stocks (2017) for this choice of tuning parameters).

Concerning implementation issues, in our experience, the tuning of the MIF algorithm (number of particles, number of iterations, etc.) can greatly affect the quality of the estimates. In particular, it seems that there is an important interplay between the tuning parameters and the initialization values of model parameters to be inferred. In comparison, our method has only one main calibration parameter in practice. In the filtering step, it is necessary to initialize the covariance matrix (i.e., T_0 in Section 2.3.2) of the state variables, conditional on the observations, but it seems that this initialization does not have a noticeable influence on the accuracy of estimates.

2.4.3 Point estimates and standard deviations for key model parameters θ

2.4.3.1 Simulation results for the first experiment ($\tau = 0$)

Three different target values for sample sizes were considered : $n = 10$, $n = 30$ and $n = 100$. Tables 2.1 and 2.2 respectively display the results for the high reporting scenario ($p = 0.8$) and the low reporting scenario ($p = 0.3$). Each table compares estimates obtained with KM and MIF. For each parameter and each estimation method, the reported values are the mean of the 500 parameter estimates, with their standard deviations in brackets.

These results show that, irrespective of the reporting rate p, when the population size N and the number of observations n per epidemic increase, the bias and the standard error of the estimates obtained decrease, whichever method is used for inference. For a given (N, n), the estimation bias is higher when the reporting rate is low ($p^* = 0.3$, where the star here designates the true value). This may be partly related to the fact that the information contained in the data decreases as p^* decreases. Both methods provide estimates with comparable levels of accuracy.

TABLE 2.1 – First experiment ($\tau = 0$). Estimation of $\theta = (\lambda, \gamma, p, i_0)$ under the constraint $s_0 + i_0 = 1$ in Setting 1 with true parameter values $(\lambda^*, \gamma^*, p^*, i_0^*)$=$(1, 1/3, 0.8, 0.01)$. For each combination of (N, n) and for each model parameter, point estimates and standard deviations are calculated as the mean of the 500 individual estimates and their standard deviations (in brackets) obtained by our Kalman-based method (KM) and Maximum Iterated Filtering (MIF). The reported values for the number of observations n correspond to the average over the 500 trajectories, with the min and max in brackets.

	$N = 1000$			$N = 2000$			$N = 10000$		
	$n = 11$	$n = 31$	$n = 101$	$n = 11$	$n = 31$	$n = 102$	$n = 10$	$n = 30$	$n = 100$
	$(7, 18)$	$(21, 51)$	$(68, 168)$	$(8, 19)$	$(23, 55)$	$(75, 179)$	$(8, 15)$	$(25, 44)$	$(81, 143)$
$\lambda^* = 1$									
KM	1.01	0.99	0.99	1.02	1.00	1.00	1.02	1.00	1.00
	(0.09)	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)	(0.03)	(0.03)	(0.03)
MIF	1.02	0.99	1.00	1.01	1.00	1.01	1.01	1.00	1.00
	(0.07)	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)	(0.02)	(0.02)	(0.02)
$\gamma^* = 1/3$									
KM	0.30	0.31	0.33	0.31	0.32	0.33	0.32	0.33	0.34
	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)	(0.02)	(0.02)	(0.02)
MIF	0.32	0.31	0.34	0.32	0.32	0.34	0.33	0.32	0.34
	(0.04)	(0.04)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
$p^* = 0.8$									
KM	0.70	0.73	0.79	0.73	0.75	0.79	0.77	0.78	0.82
	(0.10)	(0.11)	(0.06)	(0.08)	(0.11)	(0.07)	(0.05)	(0.06)	(0.05)
MIF	0.75	0.74	0.80	0.77	0.74	0.80	0.78	0.74	0.81
	(0.11)	(0.09)	(0.04)	(0.09)	(0.08)	(0.05)	(0.06)	(0.04)	(0.04)
$i_0^* = 0.01$									
KM	0.011	0.016	0.012	0.010	0.013	0.011	0.010	0.010	0.010
	(0.005)	(0.008)	(0.006)	(0.003)	(0.006)	(0.005)	(0.001)	(0.002)	(0.003)
MIF	0.011	0.012	0.011	0.011	0.012	0.011	0.010	0.011	0.010
	(0.005)	(0.004)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)

TABLE 2.2 – First experiment ($\tau = 0$). Estimation of $\theta = (\lambda, \gamma, p, i_0)$ under the constraint $s_0 + i_0 = 1$ in Setting 2 with true parameter values $(\lambda^*, \gamma^*, p^*, i_0^*)$=(1, 1/3, 0.3, 0.01). For each combination of (N, n) and for each model parameter, point estimates and standard deviations are calculated as the mean of the 500 individual estimates and their standard deviations (in brackets) obtained by KM and MIF. The reported values for the number of observations n correspond to the average over the 500 trajectories, with the min and max in brackets.

	$N = 1000$			$N = 2000$			$N = 10000$		
	$n = 11$ (7, 18)	$n = 31$ (21, 51)	$n = 101$ (68, 168)	$n = 11$ (8, 19)	$n = 31$ (23, 55)	$n = 102$ (75, 179)	$n = 10$ (8, 15)	$n = 30$ (25, 44)	$n = 100$ (81, 143)
$\lambda^* = 1$									
KM	1.01	1.04	1.00	1.00	1.02	1.01	0.99	1.02	1.00
	(0.10)	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.03)	(0.03)	(0.03)
MIF	1.02	1.07	1.01	0.99	1.03	1.02	0.98	1.01	1.00
	(0.09)	(0.07)	(0.06)	(0.06)	(0.04)	(0.05)	(0.03)	(0.02)	(0.02)
$\gamma^* = 1/3$									
KM	0.26	0.30	0.32	0.28	0.32	0.32	0.31	0.33	0.34
	(0.03)	(0.05)	(0.05)	(0.03)	(0.05)	(0.05)	(0.02)	(0.02)	(0.03)
MIF	0.27	0.30	0.31	0.28	0.32	0.32	0.31	0.34	0.33
	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
$p^* = 0.3$									
KM	0.21	0.26	0.29	0.23	0.29	0.29	0.27	0.30	0.30
	(0.03)	(0.05)	(0.05)	(0.03)	(0.05)	(0.05)	(0.02)	(0.03)	(0.03)
MIF	0.22	0.26	0.27	0.23	0.28	0.28	0.27	0.30	0.29
	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
$i_0^* = 0.01$									
KM	0.010	0.007	0.010	0.012	0.009	0.011	0.011	0.010	0.011
	(0.006)	(0.004)	(0.006)	(0.004)	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)
MIF	0.012	0.008	0.009	0.013	0.009	0.009	0.012	0.010	0.010
	(0.007)	(0.004)	(0.003)	(0.004)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)

The estimates are less computationally demanding and require less algorithmic tuning with the Kalman filtering approach. This simulation study was also performed for a second set of parameter values ($\lambda = 0.6$, $\gamma = 0.4$, $i_0 = 0.01$), under the constraint $s_0 + i_0 = 1$ and for $p = 0.8$ and $p = 0.3$, and naturally led to greater variability between simulated trajectories. These results are provided in A.5 for comparative purposes.

2.4.3.2 Simulation results for the second experiment ($\tau \neq 0$)

Here, we present the estimation results when the simulated observations are obtained with a non-zero measurement error τ, which is to be estimated. As noticed in Stocks et al. (2018), the initial conditions of the system are difficult to estimate, and usually set at plausible values. Consequently, we distinguish two situations, where either (i) i_0 is unknown and estimated ; or (ii) i_0 is known and fixed.

Unknown starting point i_0 Five different target values for sample sizes were considered : $n = 10$, $n = 30$, $n = 100$, $n = 500$, and $n = 1000$. The unknown parameters to be estimated were $\theta = (\lambda, \gamma, p, i_0, \tau)$ under the constraint $s_0 + i_0 = 1$. For the sake of clarity, we do not show the results when $N = 2000$ and $p = 0.3$. Results are displayed in Table 2.3.

TABLE 2.3 – Second experiment ($\tau \neq 0$). Estimation of $\theta = (\lambda, \gamma, p, i_0, \tau)$ under the constraint $s_0 + i_0 = 1$ in Setting 1 with true parameter values $(\lambda^*, \gamma^*, p^*, i_0^*, \tau^*)$=$(1, 1/3, 0.8, 0.01, 0.5)$. For each combination of (N, n) and for each model parameter, point estimates and standard deviations are calculated as the mean of the 500 individual estimates and their standard deviations (in brackets) obtained by our Kalman-based method and the MIF algorithm. The reported values for the number of observations n correspond to the average over the 500 trajectories, with the min and max in brackets.

	$N = 1000$					$N = 10000$				
	$n = 11$	$n = 31$	$n = 101$	$n = 501$	$n = 1001$	$n = 10$	$n = 30$	$n = 100$	$n = 499$	$n = 998$
	(7, 18)	(21, 51)	(68, 168)	(338, 833)	(676, 1665)	(8, 15)	(25, 44)	(81, 143)	(406, 716)	(811, 1430)
$\lambda^* = 1$										
KM	0.99	0.98	0.98	0.97	0.99	1.01	0.99	0.99	0.98	0.99
	(0.10)	(0.08)	(0.07)	(0.08)	(0.08)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
MIF	1.02	1.00	1.02	1.01	1.00	1.01	1.00	1.01	1.00	1.01
	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
$\gamma^* = 1/3$										
KM	0.29	0.30	0.31	0.32	0.32	0.32	0.32	0.33	0.32	0.33
	(0.03)	(0.05)	(0.05)	(0.06)	(0.07)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
MIF	0.30	0.30	0.31	0.32	0.33	0.32	0.31	0.33	0.33	0.34
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
$p^* = 0.8$										
KM	0.67	0.72	0.74	0.76	0.75	0.75	0.76	0.79	0.77	0.80
	(0.09)	(0.15)	(0.12)	(0.13)	(0.15)	(0.05)	(0.07)	(0.07)	(0.07)	(0.11)
MIF	0.70	0.70	0.74	0.75	0.78	0.75	0.72	0.78	0.77	0.82
	(0.09)	(0.09)	(0.10)	(0.08)	(0.11)	(0.05)	(0.04)	(0.05)	(0.05)	(0.06)
$i_0^* = 0.01$										
KM	0.011	0.014	0.016	0.014	0.014	0.010	0.011	0.010	0.011	0.009
	(0.005)	(0.006)	(0.006)	(0.005)	(0.010)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
MIF	0.011	0.012	0.012	0.012	0.011	0.011	0.011	0.011	0.011	0.010
	(0.004)	(0.004)	(0.003)	(0.002)	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
$\tau^* = 0.5$										
KM	0.05	0.48	0.48	0.46	0.44	0.05	0.47	0.42	0.43	0.51
	(0.16)	(0.22)	(0.13)	(0.17)	(0.21)	(0.16)	(0.19)	(0.08)	(0.09)	(0.13)
MIF	0.48	0.52	0.46	0.48	0.49	0.60	0.54	0.39	0.46	0.53
	(0.21)	(0.15)	(0.12)	(0.09)	(0.13)	(0.21)	(0.14)	(0.09)	(0.06)	(0.07)

As in the first experiment where $\tau = 0$, the results show that the estimations provided by KM and MIF are of the same order of accuracy. The pattern concerning the bias and the standard error observed in the case $\tau = 0$ also occurs when $\tau = 0.5$ is estimated, i.e., bias decreasing and accuracy increasing when N and n increase. We remark that the estimation is more difficult, inducing larger bias, when the measurement error τ is unknown, even for a quite large number of observations $n \approx 100$. Consider for example $N = 1000$, $n = 101$ and $p^* = 0.8$. The point estimate value of p obtained by KM with $\tau = 0$ (cf. Table 2.1) and $\tau \neq 0$ (cf. Table 2.3) is respectively 0.79 and 0.74. This is more marked for the second set of parameters values ($\lambda = 0.6$ and $\gamma = 0.4$), presented in A.5, which induces more variability between epidemics. For $N = 1000$, $n = 99$ and $p^* = 0.8$, comparing the results in Tables 6.1 and 6.3 shows that $\hat{p}$ passes from 0.75 to 0.66 when $\tau = 0.5$ unknown. Higher frequency observations of the epidemics lead to more satisfactory estimations : considering $n = 998$ when $\tau = 0.5$ unknown leads to $\hat{p} = 0.78$. The estimates obtained with MIF behave similarly. In summary, when the measurement error τ is non-zero and estimated, a greater number of observations is needed in order to obtain estimates without bias for both the Kalman-based and MIF methods.

Known starting point i_0 The unknown parameters to be estimated are $\theta = (\lambda, \gamma, p, \tau)$. Tables 2.4 and 2.5 respectively display the results for the high reporting scenario ($p = 0.8$) and low reporting scenario ($p = 0.3$).

TABLE 2.4 – Second experiment ($\tau \neq 0$). Estimation of $\theta = (\lambda, \gamma, p, \tau)$ with $s_0 = 0.99$ and $i_0 = 0.01$ known in Setting 1 with true parameter values $(\lambda^*, \gamma^*, p^*, \tau^*)$=$(1, 1/3, 0.8, 0.5)$. For each combination of (N, n) and for each model parameter, point estimates and standard deviations are calculated as the mean of the 500 individual estimates and their standard deviations (in brackets) obtained by KM and MIF. The reported values for the number of observations n correspond to the average over the 500 trajectories, with the min and max in brackets.

	$N = 1000$			$N = 2000$			$N = 10000$		
	$n = 11$	$n = 31$	$n = 101$	$n = 11$	$n = 31$	$n = 102$	$n = 10$	$n = 30$	$n = 100$
	$(7, 18)$	$(21, 51)$	$(68, 168)$	$(8, 19)$	$(23, 55)$	$(75, 179)$	$(8, 15)$	$(25, 44)$	$(81, 143)$
$\lambda^* = 1$									
KM	1.04	1.01	1.01	1.03	0.98	1.01	1.01	0.99	1.00
	(0.12)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.04)	(0.03)	(0.03)
MIF	1.03	1.01	1.02	1.02	1.01	1.02	1.02	1.00	1.01
	(0.08)	(0.07)	(0.07)	(0.05)	(0.05)	(0.05)	(0.02)	(0.02)	(0.02)
$\gamma^* = 1/3$									
KM	0.29	0.31	0.32	0.29	0.30	0.31	0.31	0.32	0.33
	(0.04)	(0.06)	(0.05)	(0.03)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)
MIF	0.30	0.31	0.33	0.31	0.31	0.32	0.32	0.32	0.33
	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
$p^* = 0.8$									
KM	0.69	0.76	0.76	0.71	0.71	0.75	0.74	0.76	0.79
	(0.11)	(0.16)	(0.13)	(0.09)	(0.12)	(0.10)	(0.05)	(0.06)	(0.06)
MIF	0.70	0.72	0.78	0.73	0.72	0.76	0.75	0.73	0.79
	(0.09)	(0.10)	(0.10)	(0.08)	(0.07)	(0.08)	(0.06)	(0.05)	(0.04)
$\tau^* = 0.5$									
KM	0.11	0.54	0.52	0.08	0.34	0.50	0.11	0.50	0.43
	(0.23)	(0.23)	(0.14)	(0.22)	(0.22)	(0.10)	(0.26)	(0.18)	(0.08)
MIF	0.49	0.54	0.50	0.49	0.48	0.48	0.60	0.56	0.42
	(0.22)	(0.15)	(0.11)	(0.22)	(0.14)	(0.09)	(0.23)	(0.13)	(0.07)

TABLE 2.5 – Second experiment ($\tau \neq 0$). Estimation of $\theta = (\lambda, \gamma, p, \tau)$ with $s_0 = 0.99$ and $i_0 = 0.01$ known in Setting 2 with true parameter values $(\lambda^*, \gamma^*, p^*, \tau^*)$=$(1, 1/3, 0.3, 0.5)$. For each combination of (N, n) and for each model parameter, point estimates and standard deviations are calculated as the mean of the 500 individual estimates and their standard deviations (in brackets) obtained by KM and MIF. The reported values for the number of observations n correspond to the average over the 500 trajectories, with the min and max in brackets.

	$N = 1000$			$N = 2000$			$N = 10000$		
	$n = 11$ $(7, 18)$	$n = 31$ $(21, 51)$	$n = 101$ $(68, 168)$	$n = 11$ $(8, 19)$	$n = 31$ $(23, 55)$	$n = 102$ $(75, 179)$	$n = 10$ $(8, 15)$	$n = 30$ $(25, 44)$	$n = 100$ $(81, 143)$
$\lambda^* = 1$									
KM	0.99	1.01	1.05	1.03	0.99	1.01	1.01	0.99	1.01
	(0.14)	(0.09)	(0.08)	(0.11)	(0.07)	(0.07)	(0.04)	(0.03)	(0.03)
MIF	1.05	1.06	1.05	1.06	1.02	1.03	1.02	1.01	1.01
	(0.14)	(0.08)	(0.07)	(0.10)	(0.05)	(0.05)	(0.04)	(0.02)	(0.02)
$\gamma^* = 1/3$									
KM	0.24	0.28	0.29	0.26	0.30	0.29	0.28	0.32	0.34
	(0.06)	(0.05)	(0.05)	(0.07)	(0.04)	(0.04)	(0.03)	(0.02)	(0.02)
MIF	0.23	0.29	0.29	0.24	0.30	0.30	0.28	0.32	0.34
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
$p^* = 0.3$									
KM	0.20	0.25	0.26	0.22	0.27	0.26	0.24	0.29	0.31
	(0.07)	(0.05)	(0.05)	(0.07)	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)
MIF	0.19	0.25	0.25	0.20	0.27	0.27	0.24	0.29	0.30
	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
$\tau^* = 0.5$									
KM	0.15	0.16	0.44	0.17	0.12	0.32	0.08	0.20	0.52
	(0.15)	(0.12)	(0.06)	(0.18)	(0.12)	(0.06)	(0.16)	(0.15)	(0.05)
MIF	0.41	0.26	0.44	0.45	0.24	0.36	0.50	0.30	0.50
	(0.12)	(0.10)	(0.04)	(0.12)	(0.10)	(0.06)	(0.13)	(0.09)	(0.04)

It appears that the influence of knowing or not knowing the initial condition i_0 is different according to the values of the parameters used to simulate the data. For the setting where $\lambda = 1$ and $\gamma = 1/3$, Tables 2.3 and 2.4 does not exhibit major differences between estimates. On the contrary, the impact of knowing or not knowing the initial condition i_0 is more visible when considering $\lambda = 0.6$ and $\gamma = 0.4$ (see A.5). Tables 6.3 and 6.4 show that the quality of estimates deteriorates when i_0 is unknown, leading in particular to more significant biases. For $N = 10000$, $n = 101$ and $p^* = 0.8$, $\hat{p}$ passes from 0.77 when i_0 is known to 0.65 when it is not. Once again, higher frequency observations of the epidemics lead to more satisfactory estimates (see Table 6.3). Tables 2.4 and 2.5 suggest that the estimation bias obtained for the measurement error τ increases when p^* decreases.

2.4.3.3 Additional comments

In the simulation study, we also considered cases where only the susceptible individuals are observed (not shown here). We noticed that the estimates provided by our Kalman-based method and the MIF algorithm were more accurate when considering the S rather than the I values. As the S values are several orders of magnitude larger than the I ones, a plausible explanation is that the observation noise (due to imperfect reporting and measurement errors) has a lower impact on the S values.

As for the computation times of both methods, these are sensitive to the number of observations n per epidemic : the computation time increases linearly with n. Concerning the population size

N, only the computation time for MIF-based inference increased when N increased, while our method was insensitive to it. As an example, for the scenario with $N = 10000$, $n = 30$ and $p = 0.8$ (which corresponds to Table 2.1), the average computation time for a single estimate (i.e., a single trajectory) was 31 seconds with KM and 97 seconds with the MIF algorithm. For $n = 100$, the average computation times were 81 and 147 seconds for the KM and MIF algorithms, respectively.

2.4.4 Confidence interval estimates based on profile likelihood

Following other authors (see Ionides et al. (2017) for instance), we provide profile-likelihood confidence intervals of estimated parameters, for which we briefly recall the principle. Let us denote a general parameter vector $\psi = (\psi_1, \psi_2)$, where $\psi_1 \in \mathbb{R}$ is the parameter of interest and ψ_2 contains the remaining parameters. The profile log-likelihood of ψ_1 is built by maximizing the approximate log-likelihood function (proposed in Section 2.3) over ψ_2, for fixed values of ψ_1 : $\mathcal{L}_{profile}(\psi_1) = \max_{\psi_2} \mathcal{L}(\psi_1, \psi_2)$. A 95% confidence interval for ψ_1 is given by :

$$\{\psi_1 : \mathcal{L}_{profile}(\hat{\psi}) - \mathcal{L}_{profile}(\psi_1)\} < 1.92, \tag{2.23}$$

where $\hat{\psi}$ is the maximum approximate likelihood estimator (see (2.19)). The threshold value of 1.92 comes from Wilks' theorem and corresponds to the quantile of order 0.95 of the χ^2 distribution with 1 degree of freedom.

As an illustrative example, 95% profile likelihood confidence intervals were constructed for the key epidemic parameters λ and γ on two particular trajectories of SIR simulated dynamics in the first experiment ($\tau = 0$). A graphical representation is provided in Figure 2.3 for parameter λ and in Figure 2.4 for parameter γ. The first confidence interval (left panel of both figures) is obtained with a sample of $n = 30$ observation of an SIR epidemic for a population of size $N = 2000$ with reporting rate $p = 0.3$. The second confidence interval (right panel of each Figures) is obtained with a sample of $n = 100$ observation of an SIR epidemic for a population of size $N = 10000$ with reporting rate $p = 0.8$. For each of the two parameters (playing the role of ψ_1 in (2.23)), 20 values were considered in a relevant interval containing the point estimate. For each of the 20 values of the parameter of interest, the remaining parameters (playing the role of ψ_2 in (2.23)), on which the likelihood is optimized (corresponding to $\mathcal{L}_{profile}(\psi_1)$ in (2.23))), were randomly initialized, with 10 different initialization values, the best being stored. The 20 values of maximum log-likelihood were reported on a graph, linked up by a smoothing curve. The two vertical lines, going through the intersection of this curve with the horizontal line at the y-value equal to the maximum log-likelihood for all parameters minus 1.92 (cf. equation (2.23)), determine the x-value for the CI95%. Based on Figures 2.3 and 2.4, we see that the widths of the confidence intervals CI95%$(\lambda) = [0.96, 1.10]$ and CI95%$(\gamma) = [0.31, 0.48]$ are naturally greater in the case where $N = 2000$, $n = 30$ and $p = 0.3$ (which is a more difficult case for performing estimates, due to an increased stochasticity of epidemic trajectories and significant noise in the observations) than for $N = 10000$, $n = 100$ and $p = 0.8$ (a much more tractable case with low variability amongst trajectories and low levels of noise in observations) : CI95%$(\lambda) = [0.95, 1.00]$ and CI95%$(\gamma) = [0.33, 0.36]$.

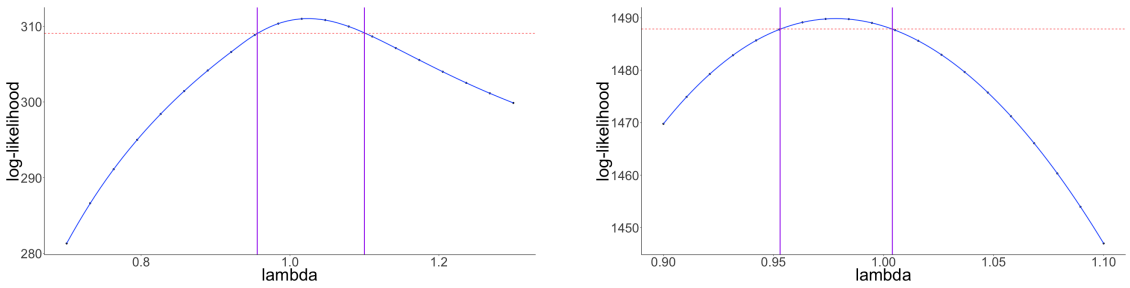

FIGURE 2.3 – Profile likelihood and confidence intervals (CI95%) for λ. Left panel : data simulated with $N = 2000$, $n = 30$ and $p = 0.3$; the true value $\lambda^* = 1$, the point estimate $\hat{\lambda} = 1.02$, and CI95% $= [0.96, 1.10]$. Right panel : data simulated with $N = 10000$, $n = 100$ and $p = 0.8$; the true value $\lambda^* = 1$, the point estimate $\hat{\lambda} = 1.00$, and CI95% $= [0.95, 1.00]$.

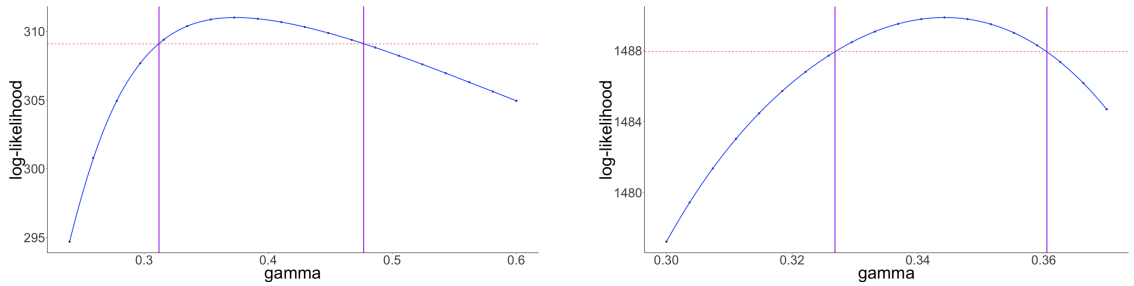

FIGURE 2.4 – Profile likelihood and confidence intervals (CI95%) for γ. Left panel : data simulated with $N = 2000$, $n = 30$ and $p = 0.3$; the true value $\gamma^* = 1/3$, the point estimate $\hat{\gamma} = 0.32$, and CI95% $= [0.31, 0.48]$. Right panel : data simulated with $N = 10000$, $n = 100$ and $p = 0.8$; the true value $\gamma^* = 1/3$, the point estimate $\hat{\gamma} = 0.34$, and CI95% $= [0.33, 0.36]$.

2.5 Application on real data

We applied our inference method on the data from an influenza outbreak that occurred in January 1978 in a boarding school in the north of England (Anonymous (1978)), with $N = 763$. The observations correspond to the daily number of infectious boys across 14 days ($n = 14$). It is known that the epidemic started from a single infectious student. Here we also assumed that the epidemic dynamics followed an SIR model. Hence, $S(0) = 762$ and $I(0) = 1$, and the parameters to be estimated are the epidemic parameters (λ, γ), the reporting rate p, and the parameter τ related to observational noise.

Estimates were performed with both KM and MIF. For the MIF method, we used the same tuning parameters values as those chosen in the simulation study. Both series of results were graphically assessed by post-predictive checks. For this, the Markov jump processes of the SIR model were simulated using each set of parameter estimates. We kept 1000 trajectories that did not exhibit early extinction, according to the theoretical criterion used in Section 2.4.1. From these 1000 trajectories, we then generated equally-spaced observations with $n = 14$. Empirical mean, 5th, 50th and 95th percentiles were extracted at each time point and superimposed on the real data (Figure 2.5).

The following estimates were obtained, with the profile likelihood-based confidence intervals (CI95%) provided in brackets :

— $\hat{\lambda}_{\text{KM}} = 1.72\ [1.61, 1.83]$; $\hat{\gamma}_{\text{KM}} = 0.48\ [0.43, 0.52]$; $\hat{p}_{\text{KM}} = 1.00\ [0.92, 1.00]$; $\hat{\tau}_{\text{KM}} = 0.91\ [0.42, 1.62]$ with KM,

— $\hat{\lambda}_{\text{MIF}} = 1.85\ [1.62, 2.15]$; $\hat{\gamma}_{\text{MIF}} = 0.47\ [0.39, 0.54]$; $\hat{p}_{\text{MIF}} = 0.97\ [0.84, 1.00]$; $\hat{\tau}_{\text{MIF}} = 1.58\ [0.80, 2.80]$ with MIF.

The estimated values for λ, γ and p are similar in both methods, but the estimated values for τ are rather different. The confidence intervals provided by the MIF method are larger than those obtained by our Kalman-based method, but this could be due to non-optimal tuning in the MIF case. Moreover, we see that the confidence interval for τ is particularly wide for both methods,

which is in agreement with the fact that a moderate number of observations is needed in order to properly estimate τ (as showed in the simulation analyses). A post-predictive check (Figure 2.5) indicates that both methods provide estimates and hence predictions that are consistent with the data. Estimation took 22.7 seconds with our method, versus 46.5 seconds using MIF.

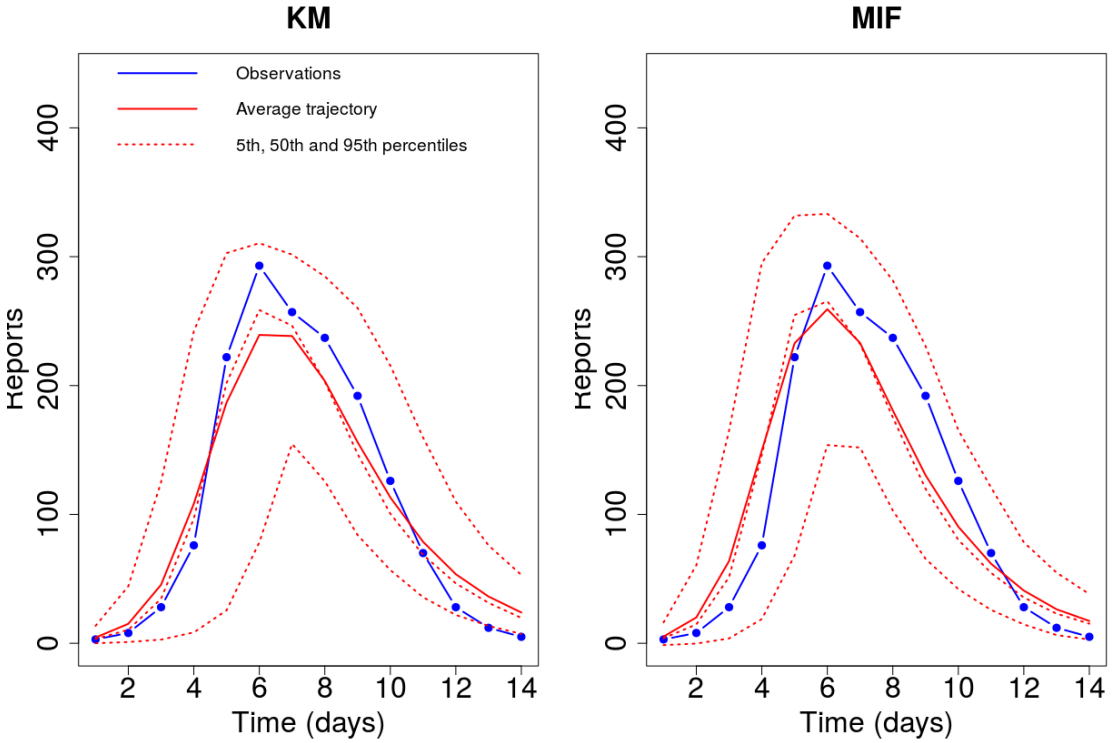

FIGURE 2.5 – Post-predictive checks for the Kalman-based (KM, left panel) and the maximum iterated filtering (MIF, right panel) estimates. In blue : observations (number of infectious boys). Solid red line : average trajectory over 1000 Markov jump processes from the estimated model. Dotted red lines : 5th, 50th, and 95th percentiles.

2.6 Discussion

In this article we have proposed a general and practical inference method for continuous-time epidemics involving discrete, partially and noisily observed time-series data. We derived a Gaussian approximation of an epidemic's density-dependent Markovian jump process underlying its dynamics using a diffusion based approach and a Gaussian approximation of observations model. This two-level Gaussian approximation allowed us to develop an inference method based on Kalman filtering for the calculation of the likelihood, to estimate key epidemic parameters (such as transmission and recovery rates), the initial state of the system (number of susceptible and infectious individuals), and parameters of the observation model (such as the reporting rate) from incomplete and noisy data (proportion of infectious individuals over time).

The performance of the estimators obtained with the Kalman-based method was investigated on simulated data under various scenarios with respect to the parameter values of epidemic and observation processes, the population size (N), the number of observations (n), and the nature of the

data (number of susceptible S or infectious I individuals over time). Performance, in terms of bias and in particular accuracy, improved when increasing N and (especially) n, and was satisfactory for a realistic observation design (e.g., $n = 30$, which corresponds in our case to one observation per day or every two days) and moderate community size ($N = 2000$).

The influence of N and n is less pronounced when data are more complete, here in the case where p, the proportion of available data—corresponding to the reporting rate—was equal to 0.8, and τ, corresponding to the measurement error, was zero. Estimation was more challenging when the measurement error τ was unknown. In the latter case, higher frequency observations were needed in order to obtain more accurate estimates. When, in addition to a non-zero measurement error, the initial point i_0 is unknown, the quality of the estimates could deteriorate in some cases.

A similar performance was observed irrespective of data type (when observations were sampled from S instead of I; results not shown). In addition, our method seemed to be little-impacted by tuning aspects. Indeed, the only obvious tuning parameter, concerning the initialization of the covariance matrix of the state variables conditionally upon the observations—in the filtering step—did not seem to influence estimation accuracy. Besides simulated data, our method provided quite plausible estimates when applied to real data from an influenza outbreak in a British boarding school in 1978, supported by the fact that the post-predictive check showed consistency with data. The good performance seen here is all the more noteworthy given that the data came with certain difficulties (low N and n).

Estimates obtained with KM were compared to those using MIF (Ionides et al. (2011), King et al. (2017)). The MIF algorithm is efficient in terms of inference quality, but computationally expensive and uses tuning parameters (number of particles, number of iterations, etc.) that are crucial for the successful functioning of the procedure. Importantly, our method does not require such specific computational calibration and its results are computed faster.

In terms of limitations of our method, we observed that the joint estimation of parameters from epidemic and observation models (λ, γ, p), along with the initial conditions of the underlying epidemic process (proportions of susceptible and infectious individuals (s_0, i_0)), can lead to difficulties when no constraint (e.g., $s_0 + i_0 = 1$) is imposed, and when only one discretized and perturbed coordinate of the system (here I) is observed. This occurred even in a "simple" scenario where $N = 10000$, $n = 100$, and $p = 0.8$ (low stochasticity and little loss of information in the data). This difficulty is no longer encountered if the two coordinates of the system (S and I) are observed. As well as this issue, two blocks of dependance between estimates were observed : (λ, γ, s_0) on the one hand, and (p, i_0) on the other. Therefore, an incorrect estimate of i_0 or s_0 will be reflected in the estimate of p and (λ, γ), respectively. One potential way to solve this problem could be to consider a prior for the initial conditions of the system. For more details on how to overcome this practical issue, see Stocks et al. (2018), Stocks (2017), who also emphasize the fact that inference algorithms are very sensitive to the initial values of the system.

Our method relies on two successive Gaussian model approximations (one for the latent state and the other for the observation model). These approximations do not seem to alter the quality of the estimates. Indeed, the small variance coefficient $N^{-1/2}$ provides an advantageous framework for the approximation of the state model, for which the Kalman filter performs very well in practice (small prediction errors). The decent accuracy of Gaussian process approximations for stochastic epidemic models has previously been highlighted (Buckingham-Jeffery et al. (2018)). Here, we went further and examined the performance of Gaussian approximations of epidemic dynamics,

not only by using a different approach based on Kalman filtering, but also by considering an even less convenient configuration where the initial conditions and observation errors had to be estimated.

Our approach can be generalized in several ways. First, although we focused in this study on the SIR model as a case study, our method is quite general since it can be extended to other mechanistic models of epidemic dynamics, including additional health states (such as an exposed state E). Second, the observations can encompass variable sampling intervals (i.e., Δ, the time step between two consecutive observations, is not necessarily constant). Third, other types of observations can be considered, both with regards to their nature (e.g., the number of new infectious individuals, which can be viewed as a function of state variables S and I) and to the error model.

Therefore, given its ease in implementation, low computation time, and satisfactory performance, we recommend the use of our Kalman filtering-based estimation method to providing an initial guess for parameters in the framework of partially observed complex epidemic dynamics.

Chapitre 3

Deuxième article : Inference in Gaussian state-space models with mixed effects for multiple epidemic dynamics

Table des matières

3.1 Introduction **67**
3.2 A mixed-effects approach for a state-space epidemic model for multiple epidemics **69**
3.2.1 The basics of the modeling framework for the case of a single epidemic 69
3.2.2 Modeling framework for multiple epidemics 72
3.3 Parametric inference **73**
3.3.1 Maximum likelihood estimation 73
3.3.2 Convergence of the SAEM-MCMC algorithm 76
3.4 Assessment of parameter estimators performances on simulated data . . . **76**
3.4.1 Simulation setting 76
3.4.2 Point estimates and standard deviations for inferred parameters 78
3.4.3 Comparison with an empirical two-step approach 80
3.5 Case study : influenza outbreaks in France **81**
3.6 Discussion **85**

Note Ce chapitre a fait l'objet d'un article soumis, (pre-print disponible, Narci et al. (2021b)).

Abstract The estimation from available data of parameters governing epidemics is a major challenge. In addition to usual issues (data often incomplete and noisy), epidemics of the same nature may be observed in several places or over different periods. The resulting possible inter-epidemic variability is rarely explicitly considered. Here, we propose to tackle multiple epidemics through a unique model incorporating a stochastic representation for each epidemic and to jointly estimate its parameters from noisy and partial observations. By building on a previous work, a Gaussian state-space model is extended to a model with mixed effects on the parameters describing simultaneously several epidemics and their observation process. An appropriate inference method is developed, by coupling the SAEM algorithm with Kalman-type filtering. Its performances are investigated on SIR simulated data. Our method outperforms an inference method separately processing each dataset. An application to SEIR influenza outbreaks in France over several years using incidence data is also carried out, by proposing a new version of the filtering algorithm. Parameter estimations highlight a non-negligible variability between influenza seasons, both in transmission and case reporting. The main contribution of our study is to rigorously and explicitly account for the inter-epidemic variability between multiple outbreaks, both from the viewpoint of modeling and inference.

Keywords Kalman filter ; Latent variables ; Parametric inference ; Random effects ; SAEM algorithm ; Stochastic compartmental models.

3.1 Introduction

Estimation from available data of model parameters describing epidemic dynamics is a major challenge in epidemiology, especially contributing to better understand the mechanisms underlying these dynamics and to provide reliable predictions. Epidemics can be recurrent over time and/or occur simultaneously in different regions. For example, influenza outbreaks in France are seasonal and can unfold in several distinct regions with different intensities at the same time. This translates into a non-negligible variability between epidemic phenomena. In practice, this inter-epidemic variability is often omitted, by not explicitly considering specific components for each entity (population, period). Instead, each data series is analysed separately and this variability is estimated empirically. Integrating in a unique model these sources of variability allows to study simultaneously the observed data sets corresponding to each spatial (e.g. region) or temporal entity (e.g. season). This approach should improve the statistical power and accuracy of the estimation of epidemic parameters as well as refine knowledge about underlying inter-epidemic variability.

An appropriate framework is represented by the mixed-effects models, which allow to describe the variability between subjects belonging to a same population from repeated data (see e.g. Pinheiro and Bates (2000), Lavielle (2014)). These models are largely used in pharmacokinetics with intra-population dynamics usually modeled by ordinary differential equations (ODE) and, in order to describe the differences between individuals, random effects on the parameters ruling these dynamics (see e.g. Collin et al. (2020)). This framework was later extended to models defined by stochastic differential equations incorporating mixed effects in the parameters of these diffusion processes (Donnet and Samson (2008), Delattre and Lavielle (2013), Donnet and Samson (2013), Delattre et al. (2018)). To our knowledge, the framework of mixed-effects models has rarely been used to analyse epidemic data, except in a very few studies. Among these, in (Prague et al. (2020)), the dynamics of the first epidemic wave of COVID-19 in France were analysed using an ODE system incorporating random parameters to take into account the variability of the dynamics between regions. Using a slightly different approach to tackle data from multiple epidemics, Bretó et al. (2020) proposed a likelihood-based inference method using particle filtering

techniques for non-linear and partially observed models. In particular, these models incorporate unit-specific parameters and shared parameters.

In addition to the specific problem of variability reflected in multiple data sets, observations of epidemic dynamics are often incomplete in various ways : only certain health states are observed (e.g. infected individuals), data are temporally discretized or aggregated, and subject to observation errors (e.g. under-reporting, diagnosis errors). Because of this incompleteness together with the non-linear structure of the epidemic models, the computation of the maximum likelihood estimator (MLE) is often not explicit. In hidden or latent variable models which are appropriate representations of incompletely observed epidemic dynamics, estimation techniques based on Expectation-Maximization (EM) algorithm can be implemented in order to compute the MLE (see e.g. Dempster et al. (1977)). However, the E-step of the EM algorithm requires that, for each parameter value θ, the conditional expectation of the complete log-likelihood given the observed data, $Q(\theta)$, can be computed. In mixed-effects models, there is generally no closed form expression for $Q(\theta)$. In such cases, this quantity can be approximated using a Monte-Carlo procedure (MCEM, Wei and Tanner (1990)), which is computationally very demanding. A more efficient alternative is the SAEM algorithm (Delyon et al. (1999)), often used in the framework of mixed-effects models (Kuhn and Lavielle (2005)), which combines at each iteration the simulation of unobserved data under the conditional distribution given the observations and a stochastic approximation procedure of $Q(\theta)$ (see also Delattre and Lavielle (2013), Donnet and Samson (2014) for the study and implementation of the SAEM algorithm for mixed-effects diffusion models).

In this article, focusing on the inference for multiple epidemic dynamics, we intend to meet two objectives. The first objective is to propose a finer modeling of multiple epidemics through a unique mixed-effects model, incorporating a stochastic representation of each epidemic. The second objective is to develop an appropriate method for jointly estimating model parameters from noisy and partial observations, able to estimate rigorously and explicitly the inter-epidemic variability. Thus, the main expected contribution is to provide accurate estimates of common and epidemic-specific parameters and to provide elements for the interpretation of the mechanisms underlying the variability between epidemics of the same nature occurring in different locations or over distinct time periods. For this purpose, we extend the Gaussian state-space model introduced in (Narci et al. (2021a)) for single epidemics to a model with mixed effects on the parameters describing simultaneously several epidemics and their observations. Then, following (Delattre and Lavielle (2013)) and building on the Kalman filtering-based inference method proposed in (Narci et al. (2021a)), we propose to couple the SAEM algorithm with Kalman-like filtering to estimate model parameters. The performances of the estimation method are investigated on simulations mimicking noisy prevalence data (*i.e.* the number of cases of disease in the population at a given time or over a given period of time). The method is then applied to the case of influenza epidemics in France over several years using noisy incidence data (*i.e.* the number of newly detected cases of the disease at a given time or over a given period of time), by proposing a new version of the filtering algorithm to handle this type of data.

The article is organized as follows. In Section 3.2 we describe the epidemic model for a single epidemic, specified for both prevalence and incidence data, and its extension to account for several epidemics through a two-level representation using the framework of mixed-effects models. Section 3.3 contains the maximum likelihood estimation method and convergence results of the SAEM algorithm. In Section 3.4, the performances of our inference method are assessed on simulated noisy prevalence data generated by SIR epidemic dynamics sampled at discrete time points. Section 3.5 is dedicated to the application case, the influenza outbreaks in France from 1990 to

2017. Section 3.6 contains a discussion and concluding remarks.

3.2 A mixed-effects approach for a state-space epidemic model for multiple epidemics

First, we sum up the approach developed in (Narci et al. (2021a)) in the case of single epidemics for prevalence data and extend it to incidence data (Section 3.2.1). By extending this approach, we propose a model for simultaneously considering several epidemics, in the framework of mixed-effects models (Section 3.2.2).

3.2.1 The basics of the modeling framework for the case of a single epidemic

The epidemic model Consider an epidemic in a closed population of size N with homogeneous mixing, whose dynamics are represented by a stochastic compartmental model with $d+1$ compartments corresponding to the successive health states of the infectious process within the population. These dynamics are described by a density-dependent Markov jump process $\mathcal{Z}(t)$ with state space $\{0,\dots,N\}^d$ and transition rates depending on a multidimensional parameter ζ. Assuming that $\mathcal{Z}(0)/N \to x_0 \neq (0,\dots,0)'$, the normalized process $\mathcal{Z}(t)/N$ representing the respective proportions of population in each health state converges, as $N \to \infty$, to a classical and well-characterized ODE :

$$\frac{\partial x}{\partial t}(\zeta,t) = b(\eta, x(\zeta,t)); \quad x(0) = x_0, \tag{3.1}$$

where $\eta = (\zeta, x_0)$ and $b(\eta,\cdot)$ is explicit and easy to derive from the Q-matrix of process $\mathcal{Z}(t)$ (see (Guy et al. (2015)), (Narci et al. (2021a))).

Two stochastic approximations of $\mathcal{Z}(t)/N$ are available : a d-dimensional diffusion process $Z(t_k)$ with drift coefficient $b(\eta,\cdot)$ and diffusion matrix $\frac{1}{N}\Sigma(\eta,\cdot)$ (which is also easily deducible from the jump functions of the density-dependent jump process, see e.g. (Narci et al. (2021a))), and a time-dependent Gaussian process $G_N(t)$ with small variance coefficient (see e.g. Britton and Pardoux (2020)), having for expression

$$G_N(t) = x(\eta,t) + \frac{1}{\sqrt{N}} g(\eta,t), \tag{3.2}$$

where $g(\eta,t)$ is a centered Gaussian process with explicit covariance matrix. There is a link between these two processes : let $W(t)$ be a Brownian motion in $\mathbb{R}^d$, then $g(\eta,t)$ is the centered Gaussian process

$$g(\eta,t) = \int_0^t \Phi(\eta,t,u)\sigma(\eta,x(\eta,u))dW(u), \quad \text{where } \sigma(\eta,x)\sigma(\eta,x)' = \Sigma(\eta,x), \tag{3.3}$$

and $\Phi(\eta,t,s)$ is the $d\times d$ resolvent matrix associated to (3.1)

$$\Phi(\eta,t,s) = \exp\left(\int_s^t \nabla_x b(\eta,x(\eta,u))\,du\right), \tag{3.4}$$

with $\nabla_x b(\eta,x)$ denoting the matrix $(\frac{\partial b_i}{\partial x_j}(\eta,x))_{1\le i,j\le d}$. In the sequel, we rely on the Gaussian process (3.2) to represent epidemic dynamics.

The epidemic is observed at discrete times $t_0 = 0 < t_1, \cdots, < t_n = T$, where n is the number of observations. Let us assume that the observation times t_k are regularly spaced, that is $t_k = k\Delta$ with

Δ the time step (but the following can be easily adapted to irregularly spaced observation times). Setting $X_k := G_N(t_k)$ and $X_0 = x_0$, the model can be written under the auto-regressive AR(1) form

$$X_k = F_k(\eta) + A_{k-1}(\eta)X_{k-1} + V_k, \quad \text{with } V_k \sim \mathcal{N}_d\left(0, T_k(\eta, \Delta)\right) \text{ and } k \geq 1. \tag{3.5}$$

All the quantities in (3.5) have explicit expressions with respect to the parameters. Indeed, using (3.1) and (3.4), we have

$$A_{k-1}(\eta) = A(\eta, t_{k-1}) = \Phi(\eta, t_k, t_{k-1}), \tag{3.6}$$

$$F_k(\eta) = F(\eta, t_k) = x(\eta, t_k) - \Phi(\eta, t_k, t_{k-1})x(\eta, t_{k-1}), \tag{3.7}$$

$$T_k(\eta, \Delta) = \frac{1}{N}\int_{t_{k-1}}^{t_k} \Phi(\eta, t_k, s)\Sigma(\eta, x(\eta, s))\ \Phi^t(\eta, t_k, s)ds. \tag{3.8}$$

Example : SIR model. As an illustrative example, we use the simple SIR epidemic model described in Figure 3.1, but other models can be considered (see e.g. the SEIR model, used in Section 3.5).

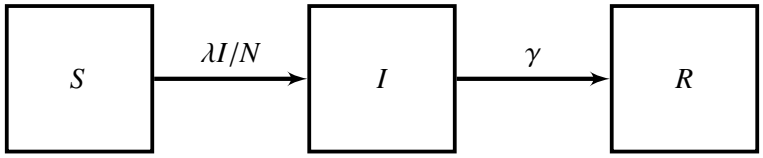

FIGURE 3.1 – SIR compartmental model with three blocks corresponding respectively to susceptible (S), infectious (I) and recovered (R) individuals. Transitions of individuals from one health state to another are governed by the transmission rate λ and the recovery rate γ, respectively.

In the SIR model, $d = 2$ and $\mathcal{Z}(t) = (S(t), I(t))'$. The parameters involved in the transition rates are λ and γ and the initial proportions of susceptible and infectious individuals are $x_0 = (s_0, i_0)'$. Denoting $\eta = (\lambda, \gamma, s_0, i_0)'$, the ODE satisfied by $x(\eta, t) = (s(\eta, t), i(\eta, t))'$ is

$$\begin{cases} \frac{\partial s}{\partial t}(\eta, t) = -\lambda s(\eta, t)i(\eta, t); & s(\eta, 0) = s_0, \\ \frac{\partial i}{\partial t}(\eta, t) = \lambda s(\eta, t)i(\eta, t) - \gamma i(\eta, t); & i(\eta, 0) = i_0. \end{cases} \tag{3.9}$$

When there is no ambiguity, we denote by s and i the solution of (3.9). Then, the functions $b(\eta, \cdot)$, $\Sigma(\eta, \cdot)$ and $\sigma(\eta, \cdot)$ are

$$b(\eta, s, i) = \begin{pmatrix} -\lambda si \\ \lambda si - \gamma i \end{pmatrix}; \quad \Sigma(\eta, s, i) = \begin{pmatrix} \lambda si & -\lambda si \\ -\lambda si & \lambda si + \gamma i \end{pmatrix}, \ \sigma(\eta, s, i) = \begin{pmatrix} \sqrt{\lambda si} & 0 \\ -\sqrt{\lambda si} & \sqrt{\gamma i} \end{pmatrix}.$$

We refer the reader to Appendix B.1 for the computation of $b(\eta, \cdot)$, $\Sigma(\eta, \cdot)$ and $\sigma(\eta, \cdot)$ in the SEIR model. Another parameterization, involving the basic reproduction number $R_0 = \frac{\lambda}{\gamma}$ and the infectious period $d = \frac{1}{\gamma}$, is more often used for SIR models. Hence, we set $\eta = (R_0, d, s_0, i_0)'$.

Observation model for prevalence data Following (Narci et al. (2021a)), we assume that observations are made at times $t_k = k\Delta, k = 1, \ldots, n$, and that some health states are not observed. The dynamics is described by the d-dimensional $AR(1)$ model detailed in (3.5). Some coordinates are not observed and various sources of noise systematically affect the observed coordinates (measurement errors, observation noises, under-reporting, etc.). This is taken into account by introducing an additional parameter μ, governing both the levels of noise and the amount of information which is available from the $q \leq d$ observed coordinates, and an operator $B(\mu) : \mathbb{R}^d \to \mathbb{R}^q$. Moreover,

we assume that, conditionally on the random variables $(B(\mu)X_k, k = 1, \dots, n)$, these noises are independent but not identically distributed. We approximate their distributions by q-dimensional Gaussian distributions with covariance matrix $P_k(\eta, \mu)$ depending on η and μ. This yields that the observations (Y_k) satisfy

$$Y_k = B(\mu)X_k + W_k, \text{ with } W_k \sim \mathcal{N}_q(0, P_k(\eta, \mu)). \tag{3.10}$$

Let us define a global parameter describing both the epidemic process and the observational process,

$$\phi = (\eta, \mu). \tag{3.11}$$

Finally, joining (3.5), (3.10) and (3.11) yields the formulation (for both epidemic dynamics and observation process) required to implement Kalman filtering methods in order to estimate the epidemic parameters :

$$\begin{cases} X_k = F_k(\eta) + A_{k-1}(\eta)X_{k-1} + V_k, \text{ with } V_k \sim \mathcal{N}_d\left(0, T_k(\eta, \Delta)\right), \ k \geq 1, \\ Y_k = B(\mu)X_k + W_k, \text{ with } W_k \sim \mathcal{N}_q\left(0, P_k(\phi)\right). \end{cases} \tag{3.12}$$

Example : SIR model (continued). The available observations could be noisy proportions of the number of infectious individuals at discrete times t_k. Denoting by p the reporting rate, one could define the operator $B(\mu) = B(p) = (0 \quad p)$ and the covariance error as $P_k(\phi) = \frac{1}{N}p(1-p)i(\eta, t_k)$ with $i(\eta, t)$ satisfying (3.9). The expression of $P_k(\phi)$ mimics the variance that would arise from assuming the observations to be obtained as binomial draws of the infectious individuals.

Observation model for incidence data For this purpose, we have extended the framework developed in (Narci et al. (2021a)). For some compartmental models, the observations (incidence) at times t_k can be written as the increments of a single or more coordinates, that is $\tilde{B}(\mu)(X_{k-1} - X_k)$ where, as above, $\tilde{B}(\mu) : \mathbb{R}^d \to \mathbb{R}^q$ is a given operator and μ are emission parameters. Let us write the epidemic model in this framework. For $k = 1, \dots, n$, let

$$\Delta_k X = X_k - X_{k-1}.$$

From (3.12), the following holds, denoting by I_d the $d \times d$ identity matrix,

$$\Delta_k X = F_k(\eta) + (A_{k-1}(\eta) - I_d)X_{k-1} + V_k. \tag{3.13}$$

As $X_{k-1} = \sum_{l=1}^{k-1} \Delta_l X + x_0$, (3.13) becomes :

$$\Delta_k X = G_k(\eta) + (A_{k-1}(\eta) - I_d) \sum_{l=1}^{k-1} \Delta_l X + V_k, \text{ with} \tag{3.14}$$

$$G_k(\eta) = x(\eta, t_k) - x_0 - \Phi(\eta, t_k, t_{k-1})(x(\eta, t_{k-1}) - x_0). \tag{3.15}$$

To model the errors that affect the data collected (Y_k), we assume that, conditionally on $(\Delta_k X, k = 1, \dots, n)$, the observations are independent and proceed to the same approximation for their distributions

$$Y_k = \tilde{B}(\mu)\Delta_k X + \tilde{W}_k; \quad \text{with } \tilde{W}_k \sim \mathcal{N}_q(0, \tilde{P}_k(\phi)). \tag{3.16}$$

Consequently, using (3.14), (3.15) and (3.16), the epidemic model for incidence data is adapted as follows :

$$\begin{cases} \Delta_k X = G_k(\eta) + (A_{k-1}(\eta) - I_d) \sum_{l=1}^{k-1} \Delta_l X + V_k, \\ Y_k = \tilde{B}(\mu)\Delta_k X + \tilde{W}_k. \end{cases} \tag{3.17}$$

Contrary to (3.5), $(\Delta_k X, k = 1, \dots, n)$ is not Markovian since it depends on all the past observations. Therefore, it does not possess the required properties of classical Kalman filtering methods. We prove in Appendix B.2 that we can propose an iterative procedure and define a new filter to compute recursively the conditional distributions describing the updating and prediction steps together with the marginal distributions of the observations from the model (3.17).

Example : SIR model (continued). Here, $\Delta_k X = \left(\frac{\Delta_k S}{N}, \frac{\Delta_k I}{N}\right)'$ and the number of new infectious individuals at times t_k is given by $\int_{t_{k-1}}^{t_k} \lambda S(t)\frac{I(t)}{N}\, dt = -\Delta_k S$. Observing a proportion p of the new infectious individuals would lead to the operator $\tilde{B}(\mu) = B(p) = (-p \quad 0)$. Mimicking binomial draws, the covariance error could be chosen as $\tilde{P}_k(\phi) = \frac{1}{N}p(1-p)(s(\eta, t_{k-1}) - s(\eta, t_k))$ where $s(\eta, t)$ satisfies (3.9).

3.2.2 Modeling framework for multiple epidemics

Consider now the situation where a same outbreak occurs in many regions or at different periods simultaneously. We use the index $1 \leq u \leq U$ to describe the quantities for each unit (e.g. region or period), where U is the total number of units. Following Section 3.2.1, for unit u, the epidemic dynamics are represented by the d-dimensional process $(X_u(t))_{t\geq 0}$ corresponding to $d+1$ infectious states (or compartments) with state space $E = [0,1]^d$. It is assumed that $(X_u(t))_{t\geq 0}$ is observed at discrete times $t_k = k\Delta$ on $[0, T_u]$, $T_u = n_u\Delta$, where Δ is a fixed time step and n_u is the number of observations, and that $Y_{u,k}$ are the observations at times t_k. Each of these dynamics has its own epidemic and observation parameters, denoted ϕ_u.

To account for intra- and inter-epidemic variability, a two level representation is considered, in the framework of mixed-effects models. First, using the discrete-time Gaussian state-space for prevalence (3.12) or for incidence data (3.17), the intra-epidemic variability is described. Second, the inter-epidemic variability is characterized by specifying a set of random parameters for each epidemic.

1. Intra-epidemic variability Let us define $X_{u,k} := X_u(t_k)$, $X_{u,0} = x_{u,0}$ and $\Delta_k X_u := X_u(t_k) - X_u(t_{k-1})$. Using (3.11), conditionally to $\phi_u = \varphi$, the epidemic observations for unit u are described as in Section 3.2.1.

For prevalence data, $1 \leq k \leq n_u$,

$$\begin{cases} X_{u,k} = F_k(\varphi) + A_{k-1}(\varphi)X_{u,k-1} + V_{u,k}, & \text{with } V_{u,k} \sim \mathcal{N}_d\left(0, T_k(\varphi, \Delta)\right), \\ Y_{u,k} = B(\varphi)X_{u,k} + W_{u,k}, & \text{with } W_{u,k} \sim \mathcal{N}_q(0, P_k(\varphi)), \end{cases} \tag{3.18}$$

(see (3.6), (3.7) and (3.8) for the expressions of $F_k(\cdot)$, $A_{k-1}(\cdot)$, $T_k(\cdot)$ and (3.10) for $B(\cdot)$ and $P_k(\cdot)$).

For incidence data,

$$\begin{cases} \Delta_k X_u = G_k(\varphi) + (A_{k-1}(\varphi) - I_d)\sum_{l=1}^{k-1}\Delta_l X_u + V_{u,k}, \\ Y_{u,k} = \tilde{B}(\varphi)\Delta_k X_u + \tilde{W}_{u,k} \quad \text{with } \tilde{W}_{u,k} \sim \mathcal{N}_q(0, \tilde{P}_k(\varphi)), \end{cases} \tag{3.19}$$

(see (3.15) for the expression of $G_k(\cdot)$ and (3.16) for $\tilde{B}(\cdot)$ and $\tilde{P}_k(\cdot)$).

2. Inter-epidemic variability We assume that the epidemic-specific parameters $(\phi_u, 1 \leq u \leq U)$ are independent and identically distributed (i.i.d) random variables with distribution defined as

follows,

$$\begin{cases} \phi_u & = h(\beta, \xi_u), \\ \xi_u & \sim \mathcal{N}_c(0, \Gamma), \end{cases} \tag{3.20}$$

where $c = \dim(\phi_u)$ and $h(\beta, x) : \mathbb{R}^c \times \mathbb{R}^c \to \mathbb{R}^c$. The vector $h(\beta, x) = (h_1(\beta, x), \ldots, h_c(\beta, x))'$ contains known link functions (a classical way to obtain parameterizations easier to handle), $\beta \in \mathbb{R}^c$ is a vector of fixed effects and $\xi_1, \ldots, \xi_U$ are random effects modeled by U i.i.d centered random variables. The fixed and random effects respectively describe the average general trend shared by all epidemics and the differences between epidemics. Note that it is sometimes possible to propose a more refined description of the inter-epidemic variability by including unit-specific covariates in (3.20). This is not considered here, without loss of generality.

Example : SIR model (continued). Let $s_{0,u} = \frac{S_u(0)}{N_u}$ and $i_{0,u} = \frac{I_u(0)}{N_u}$ where N_u is the population size in unit u. The random parameter is $\phi_u = (R_{0,u}, d_u, p_u, s_{0,u}, i_{0,u})'$ and has to fulfill the constraints

$$R_{0,u} > 1;\ d_u > 0;\ \ 0 < p_u < 1;\ 0 < s_{0,u}, i_{0,u} < 1,\ s_{0,u} + i_{0,u} \leq 1.$$

To meet these constraints, one could introduce the following function $h(\beta, x) : \mathbb{R}^5 \times \mathbb{R}^5 \to \mathbb{R}^5$:

$$\begin{cases} h_1(\beta, \xi_u) & = \exp\left[\beta_1 + \xi_{1,u}\right] + 1, \\ h_2(\beta, \xi_u) & = \exp\left[\beta_2 + \xi_{2,u}\right], \\ h_3(\beta, \xi_u) & = \frac{1}{1+\exp\left[-(\beta_3+\xi_{3,u})\right]}, \\ h_4(\beta, \xi_u) & = \frac{1}{1+\exp\left[-(\beta_4+\xi_{4,u})\right]+\exp\left[-(\beta_5+\xi_{5,u})\right]}, \\ h_5(\beta, \xi_u) & = \frac{\exp\left[-(\beta_4+\xi_{4,u})\right]}{1+\exp\left[-(\beta_4+\xi_{4,u})\right]+\exp\left[-(\beta_5+\xi_{5,u})\right]}, \end{cases} \tag{3.21}$$

where $\xi_u \sim_{i.i.d.} \mathcal{N}_5(0, \Gamma)$ and $\phi_u = h(\beta, \xi_u)$.

In this example, we supposed that all the parameters have both fixed and random effects, but it is also possible to consider a combination of random-effect parameters and purely fixed-effect parameters (see Section 3.4.1 for instance).

3.3 Parametric inference

To estimate the model parameters $\theta = (\beta, \Gamma)$, with β and Γ defined in (3.20), containing the parameters modeling the intra- and inter-epidemic variability, we develop an algorithm in the spirit of (Delattre and Lavielle (2013)) allowing to derive the maximum likelihood estimator (MLE).

3.3.1 Maximum likelihood estimation

The model introduced in Section 3.2.2 can be seen as a latent variable model with $\mathbf{y} = (y_{u,k}, 1 \leq u \leq U, 0 \leq k \leq n_u)$ the observed data and $\mathbf{\Phi} = (\phi_u, 1 \leq u \leq U)$ the latent variables. Denote respectively by $p(\mathbf{y}; \theta)$, $p(\mathbf{\Phi}; \theta)$ and $p(\mathbf{y}|\mathbf{\Phi}; \theta)$ the probability density of the observed data, of the random effects and of the observed data given the unobserved ones. By independence of the U epidemics, the likelihood of the observations $\mathbf{y}_u = (y_{u,1}, \ldots, y_{u,n_u})$ is given by :

$$p(\mathbf{y}; \theta) = \prod_{u=1}^{U} p(\mathbf{y}_u; \theta).$$

Computing the distribution $p(\mathbf{y}_u;\theta)$ of the observations for any epidemic u requires the integration of the conditional density of the data given the unknown random effects ϕ_u with respect to the density of the random parameters :

$$p(\mathbf{y}_u;\theta) = \int p(\mathbf{y}_u|\phi_u;\theta)p(\phi_u;\theta)\, d\phi_u. \tag{3.22}$$

Due to the non-linear structure of the proposed model, the integral in (3.22) is not explicit. Moreover, the computation of $p(\mathbf{y}_u|\phi_u;\theta)$ is not straightforward due to the presence of latent states in the model. Therefore, the inference algorithm needs to account for these specific features.

Let us first deal with the integration with respect to the unobserved random variables ϕ_u. In latent variable models, the use of the EM algorithm (Dempster et al. (1977)) allows to compute iteratively the MLE. Iteration k of the EM algorithm combines two steps : (1) the computation of the conditional expectation of the complete log-likelihood given the observed data and the current parameter estimate θ_k, denoted $Q(\theta|\theta_k)$ (E-step) ; (2) the update of the parameter estimates by maximization of $Q(\theta|\theta_k)$ (M-step). In our case, the E-step cannot be performed because $Q(\theta|\theta_k)$ does not have a simple analytic expression. We rather implement a Stochastic Approximation-EM (SAEM, Delyon et al. (1999)) which combines at each iteration the simulation of unobserved data under the conditional distribution given the observations (S-step) and a stochastic approximation of $Q(\theta|\theta_k)$ (SA-step).

a) General description of the SAEM algorithm Given some initial value θ_0, iteration m of the SAEM algorithm consists in the three following steps :

(S-step) Simulate a realization of the random parameters $\mathbf{\Phi}_m$ under the conditional distribution given the observations for a current parameter θ_{m-1} denoted $p(\cdot|\mathbf{y};\theta_{m-1})$.

(SA-step) Update $Q_m(\theta)$ according to

$$Q_m(\theta) = Q_{m-1}(\theta) + \alpha_m(\log p(\mathbf{y},\mathbf{\Phi}_m;\theta) - Q_{m-1}(\theta)),$$

where $(\alpha_m)_{m\geq 1}$ is a sequence of positive step-sizes s.t. $\sum_{m=1}^{\infty}\alpha_m = \infty$ and $\sum_{m=1}^{\infty}\alpha_m^2 < \infty$.

(M-step) Update the parameter estimate by maximizing $Q_m(\theta)$

$$\theta_m = \arg\max_\theta Q_m(\theta).$$

In our case, an exact sampling under $p(\cdot|\mathbf{y};\theta_{m-1})$ in the S-step is not feasible. In such intractable cases, MCMC algorithms such as Metropolis-Hastings algorithm can be used (Kuhn and Lavielle (2004)).

b) Computation of the S-step by combining the Metropoligs-Hastings algorithm with Kalman filtering techniques In the sequel, we combine the S-step of the SAEM algorithm with a MCMC procedure.

For a given parameter value θ, a single iteration of the Metropolis-Hastings algorithm consists in :

(1) Generate a candidate $\mathbf{\Phi}^{(c)} \sim q(\cdot|\mathbf{\Phi}_{m-1},\mathbf{y};\theta)$ for a given proposal distribution q

(2) Take

$$\mathbf{\Phi}_m = \begin{cases} \mathbf{\Phi}_{m-1} \text{ with probability } 1-\rho(\mathbf{\Phi}_{m-1},\mathbf{\Phi}^{(c)}), \\ \mathbf{\Phi}^{(c)} \text{ with probability } \rho(\mathbf{\Phi}_{m-1},\mathbf{\Phi}^{(c)}), \end{cases}$$

where

$$\rho(\mathbf{\Phi}_{m-1}, \mathbf{\Phi}^{(c)}) = \min\left[1, \frac{p(\mathbf{y}|\mathbf{\Phi}^{(c)};\theta)\ p(\mathbf{\Phi}^{(c)};\theta)\ q(\mathbf{\Phi}_{m-1}|\mathbf{\Phi}^{(c)},\mathbf{y};\theta)}{p(\mathbf{y}|\mathbf{\Phi}_{m-1};\theta)\ p(\mathbf{\Phi}_{m-1};\theta)\ q(\mathbf{\Phi}^{(c)}|\mathbf{\Phi}_{m-1},\mathbf{y};\theta)}\right]. \tag{3.23}$$

To compute the rate of acceptation of the Metropolis-Hastings algorithm in (3.23), we need to calculate

$$p(\mathbf{y}_u|\phi_u;\theta) = p(y_{u,0}|\phi_u;\theta)\prod_{k=1}^{n_u} p(y_{u,k}|y_{u,0},\ldots,y_{u,k-1},\phi_u;\theta),\ 1 \le u \le U.$$

Let $y_{u,k:0} := (y_{u,0},\ldots,y_{u,k})$, $k \ge 1$. In both models (3.18) and (3.19), the conditional densities $p(y_{u,k}|y_{u,k-1:0},\phi_u;\theta)$ are Gaussian densities. In model (3.18) involving prevalence data, their means and variances can be exactly computed with Kalman filtering techniques (see (Narci et al. (2021a))). In model (3.19), the Kalman filter can not be used in its standard form. We therefore develop an alternative filtering algorithm.

From now on, we omit the dependence in u and $\mathbf{\Phi}$ for sake of simplicity.

Prevalence data Let us consider model (3.12) and recall the successive steps of the filtering developed in (Narci et al. (2021a)). Assume that $X_0 \sim \mathcal{N}_d(x_0, T_0)$ and set $\hat{X}_0 = x_0, \hat{\Xi}_0 = T_0$. Then, the Kalman filter consists in recursively computing for $k \ge 1$:

1. Prediction : $\mathcal{L}(X_{k+1}|Y_k,\ldots,Y_1) = \mathcal{N}_d(\widehat{X}_{k+1}, \widehat{\Xi}_{k+1})$

$$\widehat{X}_{k+1} = F_{k+1} + A_k\overline{X}_k$$
$$\widehat{\Xi}_{k+1} = A_k\overline{T}_kA_k' + T_{k+1}$$

2. Updating : $\mathcal{L}(X_k|Y_k,\ldots,Y_1) = \mathcal{N}_d(\overline{X}_k, \overline{T}_k)$

$$\overline{X_k} = \widehat{X}_k + \widehat{\Xi}_kB'(B\widehat{\Xi}_kB' + P_k)^{-1}(Y_k - B\widehat{X}_k)$$
$$\overline{T_k} = \widehat{\Xi}_k - \widehat{\Xi}_kB'(B\widehat{\Xi}_kB' + P_k)^{-1}B\widehat{\Xi}_k$$

3. Marginal : $\mathcal{L}(Y_{k+1}|Y_k,\ldots,Y_1) = \mathcal{N}(\widehat{M}_{k+1}, \widehat{\Omega}_{k+1})$

$$\widehat{M}_{k+1} = B\widehat{X}_{k+1}$$
$$\widehat{\Omega}_{k+1} = B\widehat{\Xi}_{k+1}B' + P_{k+1}$$

Incidence data Let us consider model (3.17). Assume that $\mathcal{L}(\Delta_1X) = \mathcal{N}_d(G_1, T_1)$ and $\mathcal{L}(Y_1|\Delta_1X) = \mathcal{N}_q(\tilde{B}\Delta_1X, \tilde{P}_1)$. Let $\widehat{\Delta_1X} = G_1 = x(t_1) - x_0$ and $\widehat{\Xi}_1 = T_1$. Then, at iterations $k \ge 1$, the filtering steps are :

1. Prediction : $\mathcal{L}(\Delta_{k+1}X|Y_k,\ldots,Y_1) = \mathcal{N}_d(\widehat{\Delta_{k+1}X}, \widehat{\Xi}_{k+1})$

$$\widehat{\Delta_{k+1}X} = G_{k+1} + (A_k - I_d)\left(\sum_{l=1}^{k}\overline{\Delta_lX}\right)$$
$$\widehat{\Xi}_{k+1} = (A_k - I_d)\left(\sum_{l=1}^{k}\overline{T}_l\right)(A_k - I_d)' + T_{k+1}$$

2. Updating : $\mathcal{L}(\Delta_k X|Y_k,\dots,Y_1) = \mathcal{N}_d(\overline{\Delta_k X}, \overline{T}_k)$

$$\overline{\Delta_k X} = \widehat{\Delta_k X} + \widehat{\Xi}_k \tilde{B}'(\tilde{B}\widehat{\Xi}_k\tilde{B}' + \tilde{P}_k)^{-1}(Y_k - \tilde{B}\widehat{\Delta_k X})$$
$$\overline{T}_k = \widehat{\Xi}_k - \widehat{\Xi}_k\tilde{B}'(\tilde{B}\widehat{\Xi}_k\tilde{B}' + \tilde{P}_k)^{-1}\tilde{B}\widehat{\Xi}_k$$

3. Marginal : $\mathcal{L}(Y_{k+1}|Y_k,\dots,Y_1) = \mathcal{N}(\widehat{M}_{k+1}, \widehat{\Omega}_{k+1})$

$$\widehat{M}_{k+1} = \tilde{B}\widehat{\Delta_{k+1}X}$$
$$\widehat{\Omega}_{k+1} = \tilde{B}\widehat{\Xi}_{k+1}\tilde{B}' + \tilde{P}_{k+1}$$

The equations are deduced in Appendix B.2, the difficult point lying in the prediction step, *i.e.* the derivation of the conditional distribution $\mathcal{L}(\Delta_{k+1}X|Y_k,\cdots,Y_1)$.

3.3.2 Convergence of the SAEM-MCMC algorithm

Generic assumptions guaranteeing the convergence of the SAEM-MCMC algorithm were stated in (Kuhn and Lavielle (2004)). These assumptions mainly concern the regularity of the model (see assumptions **(M1-M5)**) and the properties of the MCMC procedure used in step S (**SAEM3'**). Under these assumptions, and providing that the step sizes (α_m) are such that $\sum_{m=1}^{\infty}\alpha_m = \infty$ and $\sum_{m=1}^{\infty}\alpha_m^2 < \infty$, then the sequence (θ_m) obtained through the iterations of the SAEM-MCMC algorithm converges almost surely toward a stationary point of the observed likelihood.

Let us remark that by specifying the inter-epidemic variability through the modeling framework of Section 3.2.2, our approach for multiple epidemics fulfills the exponentiality condition stated in **(M1)** provided that all the components of ϕ_u are random. Hence the algorithm proposed above converges almost surely toward a stationary point of the observed likelihood under the standard regularity conditions stated in **(M2-M5)** and assumption **(SAEM3')**.

3.4 Assessment of parameter estimators performances on simulated data

First, the performances of our inference method are assessed on simulated stochastic SIR dynamics. Second, the estimation results are compared with those obtained by an empirical two-step approach.

For a given population of size N and given parameter values, we use the Gillespie algorithm (Gillespie (1977)) to simulate a two-dimensional Markov jump process $\mathcal{Z}(t) = (S(t), I(t))'$. Then, choosing a sampling interval Δ and a reporting rate p, we consider prevalence data $(O(t_k), k = 1,\dots,n)$ simulated as binomial trials from a single coordinate of the system $I(t_k)$.

3.4.1 Simulation setting

Model Recall that the epidemic-specific parameters are $\phi_u = (R_{0,u}, d_u, p_u, s_{0,u}, i_{0,u})'$. In the sequel, for all $u \in \{1,\dots,U\}$, we assume that $R_{0,u} > 1$ and $0 < p_u < 1$ are random parameters. We also set $s_{0,u} + i_{0,u} = 1$ (which means that the initial number of recovered individuals is zero), with $0 < i_{0,u} < 1$ being a random parameter. Moreover, we consider that the infectious period $d_u = d > 0$ is a fixed parameter since the duration of the infectious period can reasonably be assumed constant between different epidemics. It is important to note that the case study is outside the scope of the exponential model since a fixed parameter has been included. We refer the reader

to Appendix B.3 for implementation details.

Four fixed effects $\beta \in \mathbb{R}^4$ and three random effects $\xi_u = (\xi_{1,u}, \xi_{3,u}, \xi_{4,u})' \sim \mathcal{N}_3(0, \Gamma)$ are considered. Therefore, using (3.20) and (3.21), we assume the following model for the fixed and random parameters :

$$\phi_u = (R_{0,u}, d_u, p_u, i_{0,u})' = h(\beta, \xi_u), \qquad \text{with} \tag{3.24}$$

$$\begin{aligned} h_1(\beta, \xi_u) &= \exp\left[\beta_1 + \xi_{1,u}\right] + 1, \\ h_2(\beta, \xi_u) &= \exp\left[\beta_2\right], \\ h_i(\beta, \xi_u) &= \frac{1}{1 + \exp\left[-(\beta_i + \xi_{i,u})\right]}, \ i = 3, 4. \end{aligned}$$

In other words, random effects on (R_0, p, i_0) and fixed effect on d are considered. Moreover, these random effects come from a priori independent sources, so that there is no reason to consider correlations between $\xi_{1,u}$, $\xi_{3,u}$ and $\xi_{4,u}$, and we can assume in this set-up a diagonal form for the covariance matrix $\Gamma = \operatorname{diag} \Gamma_i$, $i \in \{1, 3, 4\}$.

Parameter values We consider two settings (denoted respectively (i) and (ii) below) corresponding to two levels of inter-epidemic variability (resp. high and moderate). The fixed effects values β are chosen such that the intrinsic stochasticity of the epidemic dynamics is significant (a second set of fixed effects values leading to a lower intrinsic stochasticity is also considered ; see Appendix B.4 for details).

— Setting (i) : $\beta = (-0.81, 0.92, 1.45, -2.20)'$ and $\Gamma = \operatorname{diag}(0.47^2, 1.50^2, 0.75^2)$ corresponding to $\mathbb{E}(R_{0,u}) = 1.5$, $CV_{R_{0,u}} = 17\%$; $d = 2.5$; $\mathbb{E}(p_u) \approx 0.74$, $CV_{p_u} \approx 31\%$; $\mathbb{E}(i_{0,u}) \approx 0.12$, $CV_{i_{0,u}} \approx 66\%$;

— Setting (ii) : $\beta = (-0.72, 0.92, 1.45, -2.20)'$ and $\Gamma = \operatorname{diag}(0.25^2, 0.90^2, 0.50^2)$ corresponding to $\mathbb{E}(R_{0,u}) = 1.5$, $CV_{R_{0,u}} = 8\%$; $d = 2.5$; $\mathbb{E}(p_u) \approx 0.78$, $CV_{p_u} \approx 18\%$; $\mathbb{E}(i_{0,u}) \approx 0.11$, $CV_{i_0} \approx 45\%$;

where CV_ϕ stands for the coefficient of variation of a random variable ϕ. Let us note that the link between ϕ_u and (β, ξ_u) for p and i_0 does not have an explicit expression.

Data simulation The population size is fixed to $N_u = N = 10,000$. For each $U \in \{20, 50, 100\}$, $J = 100$ data sets, each composed of U SIR epidemic trajectories, are simulated. Independent samplings of $\left(\phi_{u,j} = (R_{0,u}, d_u, p_u, i_{0,u})'_j\right)$, $u = 1, \dots U$, $j = 1, \dots, J$, are first drawn according to model (3.24). Then, conditionally to each parameter set $\phi_{u,j}$, a bidimensionnal Markov jump process $\mathcal{Z}_{u,j}(t) = (S_{u,j}(t), I_{u,j}(t))'$ is simulated. Normalizing $\mathcal{Z}_{u,j}(t)$ with respect to N_u and extracting the values of the normalized process at regular time points $t_k = k\Delta$, $k = 1, \dots, n_{u,j}$, gives the $X_{u,k,j} = \left(\frac{S_{u,k,j}}{N_u}, \frac{I_{u,k,j}}{N_u}\right)'$'s. A fixed discretization time step is used, *i.e.* the same value of Δ is used to simulate all the epidemic data. For each epidemic, $T_{u,j}$ is defined as the first time point at which the number of infected individuals becomes zero. Two values of Δ are considered ($\Delta \in \{0.425, 2\}$) corresponding to an average number of time-point observations $\overline{n}_j = \frac{1}{U}\sum_{u=1}^{U} n_{u,j} \in \{20, 100\}$. Only trajectories that did not exhibit early extinction were considered for inference. The theoretical proportion of these trajectories is given by $1 - (1/R_0)^{I_0}$ (Andersson and Britton (2000)). Then, given the simulated $X_{u,k,j}$'s and parameters $\phi_{u,j}$'s, the observations $Y_{u,k,j}$ are generated from binomial distributions $\mathcal{B}(I_{u,k,j}, p_{u,j})$.

3.4.2 Point estimates and standard deviations for inferred parameters

Tables 3.1 and 3.2 show the estimates of the expectation and standard deviation of the mixed effects ϕ_u, computed from the estimations of β and Γ using functions h defined in (3.24), for settings (i) and (ii). For each parameter, the reported values are the mean of the $J = 100$ parameter estimates $\phi_{u,j}$, $j \in \{1, \ldots, J\}$, and their standard deviations in brackets.

TABLE 3.1 – Estimates for setting (i) : high inter-epidemic variability. For each combination of $(\bar{n}, U)$ and for each model parameter (defined in the first line of the table), point estimates and precision are calculated as the mean of the $J = 100$ individual estimates and their standard deviations (in brackets).

Parameters		$\mathbb{E}(R_{0,u})$	d	$\mathbb{E}(p_u)$	$\mathbb{E}(i_{0,u})$	$\mathrm{sd}(R_{0,u})$	$\mathrm{sd}(p_u)$	$\mathrm{sd}(i_{0,u})$
True values		**1.500**	**2.500**	**0.739**	**0.119**	**0.250**	**0.226**	**0.079**
$\bar{n} = 20$	$U = 20$	1.580	2.584	0.688	0.126	0.335	0.193	0.078
		(*0.135*)	(*0.293*)	(*0.117*)	(*0.024*)	(*0.151*)	(*0.051*)	(*0.020*)
	$U = 50$	1.574	2.538	0.704	0.122	0.359	0.201	0.079
		(*0.111*)	(*0.220*)	(*0.089*)	(*0.019*)	(*0.149*)	(*0.030*)	(*0.014*)
	$U = 100$	1.583	2.564	0.700	0.124	0.385	0.199	0.081
		(*0.105*)	(*0.210*)	(*0.083*)	(*0.015*)	(*0.134*)	(*0.023*)	(*0.011*)
$\bar{n} = 100$	$U = 20$	1.501	2.502	0.734	0.118	0.292	0.217	0.075
		(*0.080*)	(*0.159*)	(*0.059*)	(*0.021*)	(*0.105*)	(*0.035*)	(*0.019*)
	$U = 50$	1.510	2.522	0.729	0.120	0.305	0.217	0.080
		(*0.054*)	(*0.126*)	(*0.038*)	(*0.014*)	(*0.070*)	(*0.022*)	(*0.012*)
	$U = 100$	1.503	2.508	0.738	0.119	0.308	0.216	0.079
		(*0.047*)	(*0.097*)	(*0.030*)	(*0.010*)	(*0.054*)	(*0.016*)	(*0.009*)

The results show that all the point estimates are close to the true values (relatively small bias), whatever the inter-epidemic variability setting, even for small values of $\bar{n}$ and U. When the number of epidemics U increases, the standard error of the estimates decreases, but it does not seem to have a real impact on the estimation bias. Besides, observations of higher frequency of the epidemics (large $\bar{n}$) lead to lower bias and standard deviations. It is particularly marked concerning both expectation and standard deviations of the random parameters $R_{0,u}$ and p_u. Irrespective to the level of inter-epidemic variability, the estimations are quite satisfactory. While standard deviations of $R_{0,u}$ are slightly over-estimated, even for large U and $\bar{n}$, this trend in bias does not affect the standard deviations of p_u and $i_{0,u}$.

For a given data set, Figure 3.2 displays convergence graphs of the SAEM algorithm for each estimates of model parameters in setting (i) with $U = 100$ and $\bar{n} = 100$. Although the model does not belong to the curved exponential family, convergence of model parameters towards their true value is obtained for all parameters.

TABLE 3.2 – Estimates for setting (ii) : moderate inter-epidemic variability. For each combination of $(\bar{n}, U)$ and for each model parameter (defined in the first line of the table), point estimates and precision are calculated as the mean of the $J = 100$ individual estimates and their standard deviations (in brackets).

Parameters		$\mathbb{E}(R_{0,u})$	d	$\mathbb{E}(p_u)$	$\mathbb{E}(i_{0,u})$	$\mathrm{sd}(R_{0,u})$	$\mathrm{sd}(p_u)$	$\mathrm{sd}(i_{0,u})$
True values		**1.500**	**2.500**	**0.777**	**0.109**	**0.125**	**0.143**	**0.049**
$\bar{n} = 20$	$U = 20$	1.619 (*0.120*)	2.764 (*0.256*)	0.666 (*0.099*)	0.127 (*0.022*)	0.190 (*0.106*)	0.117 (*0.034*)	0.053 (*0.014*)
	$U = 50$	1.638 (*0.103*)	2.789 (*0.233*)	0.653 (*0.087*)	0.128 (*0.018*)	0.213 (*0.099*)	0.122 (*0.018*)	0.056 (*0.010*)
	$U = 100$	1.623 (*0.081*)	2.769 (*0.194*)	0.658 (*0.075*)	0.128 (*0.013*)	0.209 (*0.056*)	0.122 (*0.017*)	0.056 (*0.007*)
$\bar{n} = 100$	$U = 20$	1.540 (*0.066*)	2.627 (*0.143*)	0.732 (*0.057*)	0.118 (*0.017*)	0.176 (*0.055*)	0.143 (*0.035*)	0.050 (*0.012*)
	$U = 50$	1.539 (*0.044*)	2.622 (*0.098*)	0.733 (*0.041*)	0.117 (*0.009*)	0.183 (*0.038*)	0.145 (*0.018*)	0.052 (*0.007*)
	$U = 100$	1.541 (*0.040*)	2.629 (*0.078*)	0.732 (*0.030*)	0.118 (*0.008*)	0.187 (*0.035*)	0.149 (*0.016*)	0.053 (*0.006*)

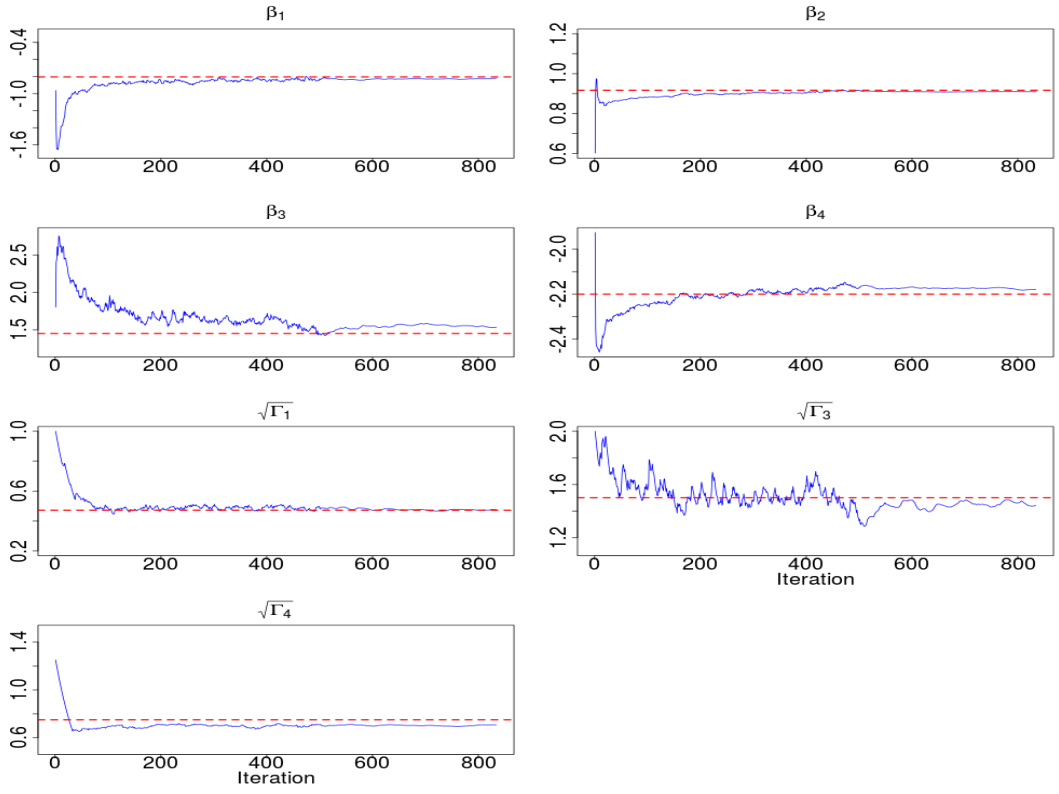

FIGURE 3.2 – Convergence graphs of the SAEM algorithm for estimates of $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$ and $\mathrm{diag}(\Gamma) = (\Gamma_1, \Gamma_3, \Gamma_4)$. Setting (i) with $U = 100$ and $\bar{n} = 100$. Parameter values at each iteration of the SAEM algorithm (plain blue line) and true values of model parameters (dotted red line).

3.4.3 Comparison with an empirical two-step approach

The inference proposed method (referred to as SAEM-KM) is compared to an empirical two-step approach not taking into account explicitly mixed effects in the model. For that purpose, let us consider the method presented in (Narci et al. (2021a)) (referred to as KM) performed in two steps : first, we compute the estimates $\hat{\phi}_u$ independently on each of the U trajectories. Second, the empirical mean and variance of the $\hat{\phi}_u$'s are computed. We refer the reader to Appendix B.3 for practical considerations on implementation of the KM method.

Let us consider $\bar{n} = 50$ and $U \in \{20, 100\}$. Figure 3.3 displays the distribution of the bias of the parameter estimates $\phi_{u,j}$, $j \in \{1, \ldots, J\}$, $J = 100$, obtained with SAEM-KM and KM for simulation settings (i) and (ii).

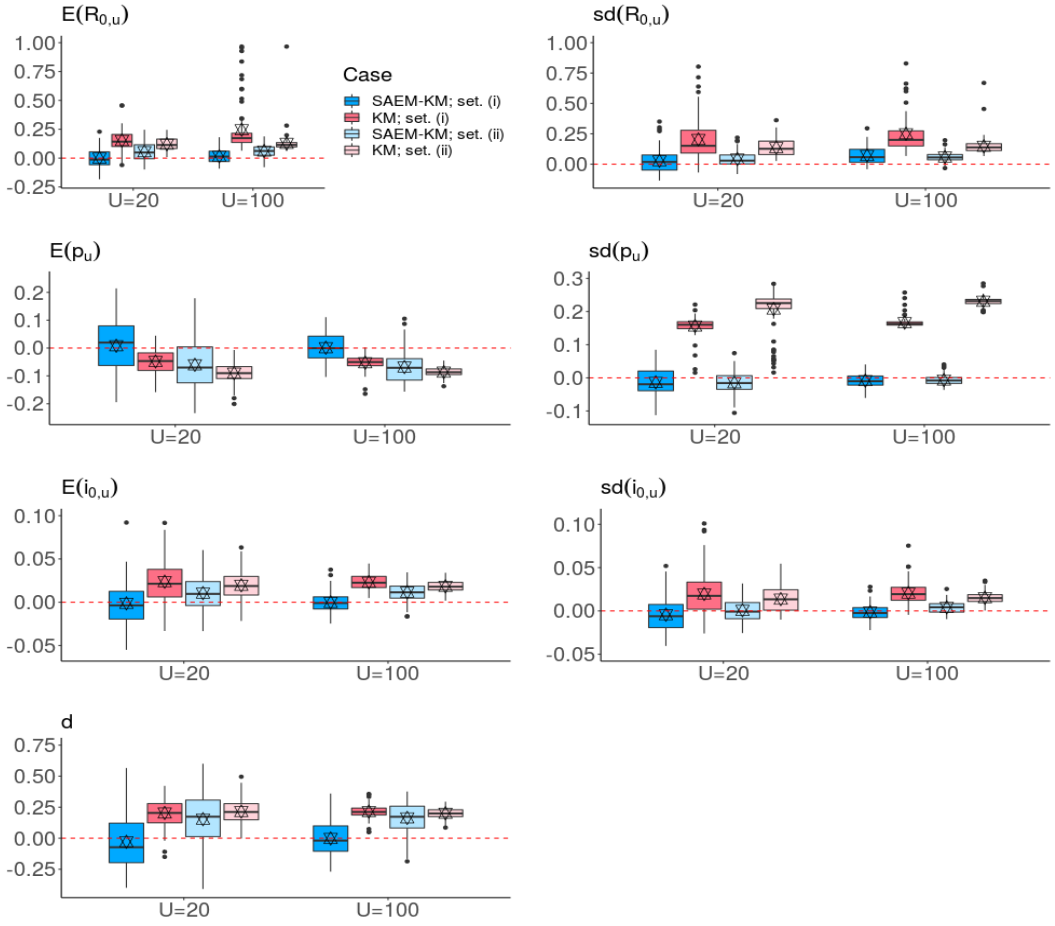

FIGURE 3.3 – Boxplots (25th, 50th and 75th percentiles) of the bias of the estimates of each model parameter, with $\bar{n} = 50$, obtained with SAEM-KM (blue boxes) and KM (red boxes). Two levels : $U = 20$ and $U = 100$ epidemics. Dark colours : high inter-epidemic variability (setting (i)). Light colours : moderate inter-epidemic variability (setting (ii)). The symbol represents the estimated mean bias. For sake of clarity, we removed extreme values from the graphical representation. This concerns only the parameter R_0 and the KM method : 37 values for $\mathbb{E}(R_{0,u})$ (35 in setting (i), 2 in setting (ii)) and 50 values for $\text{sd}(R_{0,u})$ (47 in setting (i), 3 in setting (ii)).

We notice a clear advantage to consider the mixed-effects structure. Overall, the results show that

SAEM-KM outperforms KM. This is more pronounced for standard deviation estimates in the large inter-epidemic variability setting (i) than in the moderate inter-epidemic variability setting (ii). Concerning the expectation estimates, their dispersion around the median is lower for KM than for SAEM-KM, especially in setting (ii), but the bias of KM estimates is also higher. When the inter-epidemic variability is high (setting (i)), the performances of the two inference methods are substantially different. In particular, KM sometimes fails to provide plausible estimates (especially for parameter R_0).

We also tested other values for $\bar{n}$ and N (not shown here), e.g. $\bar{n} = 20$ (lower amount of information) and $N = 2000$ (higher intrinsic variability of epidemics). In such cases, KM also failed to provide satisfying estimations whereas the mixed-effects approach was much more robust.

3.5 Case study : influenza outbreaks in France

Data The SAEM-KM method is evaluated on a real data set of influenza outbreaks in France provided by the Réseau Sentinelles (url : `www.sentiweb.fr`). We use the daily number of influenza-like illness (ILI) cases between 1990-2017, considered as a good proxy of the number of new infectious individuals. The daily incidence rate was expressed per 100, 000 inhabitants. To select epidemic periods, we chose the arbitrary threshold of weekly incidence of 160 cases per 100, 000 inhabitants (Cauchemez et al. (2008)), leading to 28 epidemic dynamics. Two epidemics have been discarded due to their bimodality (1991-1992 and 1998). Therefore, $U = 26$ epidemic dynamics are considered for inference.

Compartmental model Let us consider the SEIR model (see Figure 3.4). An individual is considered exposed (E) when infected but not infectious. Denote $\eta = (\lambda, \epsilon, \gamma, x_0)$, with $x_0 = (s_0, e_0, i_0, r_0)$, the parameters involved in the transition rates, where ϵ is the transition rate from E to I. ODEs of the SEIR model are as follows :

$$\begin{cases} \frac{ds}{dt}(\eta, t) &= -\lambda s(\eta, t) i(\eta, t), \\ \frac{de}{dt}(\eta, t) &= \lambda s(\eta, t) i(\eta, t) - \epsilon e(\eta, t), \\ \frac{di}{dt}(\eta, t) &= \epsilon e(\eta, t) - \gamma i(\eta, t), \\ \frac{dr}{dt}(\eta, t) &= \gamma i(\eta, t), \\ x_0 &= (s_0, e_0, i_0, r_0). \end{cases} \tag{3.25}$$

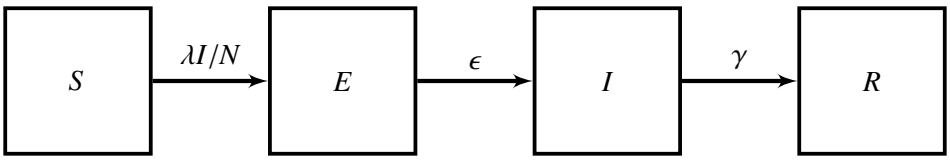

FIGURE 3.4 – SEIR compartmental model with four blocks corresponding respectively to susceptible (S), exposed (E), infectious (I) and recovered (R) individuals. Transitions of individuals from one health state to another are governed by the transmission rate λ, the incubation rate ϵ and the recovery rate γ.

Another parametrization exhibits the basic reproduction number $R_0 = \frac{\lambda}{\gamma}$, the incubation period $d_E = \frac{1}{\epsilon}$ and the infectious period $d_I = \frac{1}{\gamma}$. Thus, the epidemic pamareters are $\eta = (R_0, d_E, d_I, s_0, e_0, i_0)'$. Let us describe the two-layer model used in the sequel.

Intra-epidemic variability For each epidemic u, let $X_u = \left(\frac{S_u}{N_u}, \frac{E_u}{N_u}, \frac{I_u}{N_u}\right)'$ and

$$\eta_u = (R_{0,u}, d_{E,u}, d_{I,u}, s_{u,0}, e_{u,0}, i_{u,0}),$$

where the population size is fixed at $N_u = N = 100,000$. Denote by $\text{Inc}_u(t_k)$ the number of newly infected individuals at time t_k for epidemic u. We have

$$\text{Inc}_u(t_k) = \int_{t_{k-1}}^{t_k} \frac{1}{d_{E,u}} E_u(t)dt = S_u(t_{k-1}) - S_u(t_k) + E_u(t_{k-1}) - E_u(t_k) = -(\Delta_k S_u + \Delta_k E_u).$$

Observations are modeled as incidence data observed with Gaussian noises. We draw our inspiration from (Bretó (2018)) to account for over-dispersion in data. Therefore, assuming a reporting rate p_u for epidemic u, the mean and the variance of the observed newly infected individuals are respectively defined as $p_u\text{Inc}_u(t_k)$ and $p_u\text{Inc}_u(t_k) + \tau_u^2 p_u^2 \text{Inc}_u(t_k)^2$, where parameter τ_u is introduced to handle over-dispersion in the data. Denote $\phi_u = \left(\eta_u, p_u, \tau_u^2\right)$. Therefore, we use the model defined in (3.19) with $\Delta_k X_u = \left(\frac{\Delta_k S_u}{N}, \frac{\Delta_k E_u}{N}, \frac{\Delta_k I_u}{N}\right)'$, $V_{u,k} \sim \mathcal{N}_d(0, T_k(\phi_u, \Delta))$, $\tilde{W}_{u,k} \sim \mathcal{N}_q(0, \tilde{P}_k(\phi_u))$, $G_k(\cdot)$, $A_{k-1}(\cdot)$ and $T_k(\cdot)$ deriving from (6.2) in Appendix B.1, $\tilde{B}(\phi_u) = (-p_u \quad -p_u \quad 0)$ and

$$\tilde{P}_k(\phi_u) = \frac{1}{N}\tilde{B}(\phi_u)\Delta_k x_u + \tau_u^2 \left(\tilde{B}(\phi_u)\Delta_k x_u\right)^2,$$

where $x(\cdot, t)$ is the ODE solution of (3.25).

Inter-epidemic variability Let us first comment on the duration of the incubation period d_E and of the infectious period d_I. Studies in the literature found discrepant values of these durations (see Cori et al. (2012) for a review), varying from 0.64 (Fraser et al. (2009)) to 3.0 (Pourbohloul et al. (2009)) days for the incubation period and from 1.27 (Fraser et al. (2009)) to 8.0 (Pourbohloul et al. (2009)) days for the infectious period. For example, Cori et al. (2012) estimated that $d_E = 1.6$ and $d_I = 1.0$ days on average using excretion profiles from experimental infections. In two other papers, these durations were fixed according to previous studies (e.g. Mills et al. (2004), Ferguson et al. (2005)) : $(d_E, d_I) = (1.9, 4.1)$ days (Chowell et al. (2008)) ; $(d_E, d_I) = (0.8, 1.8)$ days (Bague-lin et al. (2013)). Performing a systematic review procedure from viral shedding and/or symptoms, Carrat et al. (2008) estimated d_E to be between 1.7 and 2.0 on average. For identifiability reasons, we consider the latent and infectious periods d_E and d_I known and test three combinations of values : $(d_E, d_I) = (1.6, 1.0)$, $(0.8, 1.8)$ and $(1.9, 4.1)$.

We consider that the basic reproduction number R_0 and the reporting rate p are random, reflecting the assumptions that the transmission rate of the pathogen varies from season to season and the reporting could change over the years. Moreover, we assume $e_u(0) = i_u(0)$ random and unknown (*i.e.* the proportion of initial exposed and infectious individuals is variable between epidemics). Cauchemez et al. (2008) assumed that at the start of each influenza season, a fixed average of 27% of the population is immune, that is $r_{0,u} = r_0 = 0.27$. To assess the robustness of the model with respect to the r_0 value, we test three values : $r_0 \in \{0.1, 0.27, 0.5\}$. This leads to $s_{0,u} = 1 - r_0 - 2i_{0,u}$ random and unknown. Finally, we assume that $\tau_u^2 = \tau^2$ is fixed and unknown. To sum up, we have to consider in the model : known parameters $(d_E, d_I) \in \{(0.8, 1.8), (1.6, 1.0), (1.9, 4.1)\}$ and $r_0 \in \{0.1, 0.27, 0.5\}$; fixed and unknown parameter τ^2 ; random and unknown parameters R_0, i_0 and p.

Therefore, using (3.20), we consider the following model for random parameters :

$$\phi_u = \left(R_{0,u}, p_u, i_{0,u}, \tau^2\right)' = h(\beta, \xi_u), \qquad \text{with}$$

$$
\begin{aligned}
h_1(\beta, \xi_u) &= \exp\left[\beta_1 + \xi_{1,u}\right] + 1, \\
h_j(\beta, \xi_u) &= \frac{1}{1 + \exp\left[-(\beta_j + \xi_{j,u})\right]}, \ j = 2, 3, \\
h_4(\beta, \xi_u) &= \exp\left[\beta_4\right],
\end{aligned}
$$

where fixed effects $\beta \in \mathbb{R}^4$ and the random effects are $\xi_u \sim_{i.i.d.} \mathcal{N}_3(0, \Gamma)$ with Γ a covariance matrix assumed to be diagonal.

Parameter estimates We consider nine models with different combinations of values of $((d_E, d_I), r_0)$. Using importance sampling techniques, we estimate the observed log-likelihood of each model from the estimated parameters values initially obtained with the SAEM algorithm. Table 3.3 provides the estimated log-likelihood values of the nine models of interest. Irrespectively of the r_0 value, we find that the model with $(d_E, d_I) = (1.9, 4.1)$ outperforms the two other models in terms of log-likelihood value. Moreover, for a given combination of values of (d_E, d_I), the estimated log-likelihood values are quite similar according to the three r_0 tested values.

TABLE 3.3 – Estimated values of the observed log-likelihood of the model obtained by testing nine combinations of values of $((d_E, d_I), r_0)$.

(d_E, d_I)	r_0	Estimated log-likelihood
(0.8,1.8)	0.1	9011.752
	0.27	8827.870
	0.5	8499.452
(1.6,1.0)	0.1	9147.108
	0.27	8961.991
	0.5	8643.562
(1.9,4.1)	0.1	10270.000
	0.27	10216.260
	0.5	9905.436

Let us focus on the model with $(d_E, d_I) = (1.9, 4.1)$. Table 3.4 presents the estimation results of the model parameters obtained by testing the three values of r_0 : 0.1, 0.27 and 0.5.

TABLE 3.4 – Estimates of the mean, 5th and 95th percentiles and coefficient of variation (CV) for model parameters $\left(R_{0,u}, i_{0,u}, p_u, \tau^2\right)'$, assuming $(d_E, d_I) = (1.9, 4.1)$ and testing three values of r_0 : 0.1, 0.27 and 0.5. For fixed parameter, only the estimated mean is available.

		$R_{0,u}$	p_u	$i_{0,u}$	τ^2
Estimated mean	$r_0 = 0.1$	1.810	0.069	0.010	0.025
	$r_0 = 0.27$	2.238	0.084	0.008	0.013
	$r_0 = 0.5$	3.281	0.119	0.006	0.037
Estimated [5th,95th] percentiles	$r_0 = 0.1$	[1.470,2.264]	[0.026,0.138]	[0.003,0.023]	—
	$r_0 = 0.27$	[1.787,2.825]	[0.031,0.169]	[0.002,0.019]	—
	$r_0 = 0.5$	[2.696,3.977]	[0.044,0.238]	[0.002,0.014]	—
Estimated CV	$r_0 = 0.1$	14 %	53 %	67 %	—
	$r_0 = 0.27$	14 %	52 %	72 %	—
	$r_0 = 0.5$	12 %	51 %	74 %	—

The average estimated value of R_0 is quite contrasted according to the r_0 value : between 1.81 and 3.28 from $r_0 = 0.1$ to $r_0 = 0.5$. By comparison, in (Cauchemez et al. (2008)), R_0 is estimated to be 1.7 during school term, and 1.4 in holidays, using a population structured into households and schools. Chowell et al. (2008) estimated a different reproduction number $\tilde{R} = (1 - r_0)R_0 = 1.3$, measuring the transmissibility at the beginning of an epidemic in a partially immune population, from mortality data. In our case, the average value of $\tilde{R}$ is estimated to 1.63, 1.63 and 1.64 when $r_0 = 0.1$, 0.27 and 0.5 respectively. Therefore, given the nature of the observations (new infected individuals) and the considered model, this appears to be difficult to correctly identify R_0 together with r_0. Indeed, the fraction of immunized individuals at the beginning of each seasonal influenza epidemic is an important parameter for the epidemic dynamics, but its value is not well known. This has implications for the stability of the estimation of the other parameters. Interestingly, the average reporting rate is estimated particularly low (around 10% irrespective of the r_0 value). Moreover, we observe that R_0 together with p and i_0 seem to be variable from season to season, with moderate coefficient of variation CV($R_{0,u}$) close to 15% and high coefficients of variation CV(p_u) and CV($i_{0,u}$) around 50% and 70% respectively.

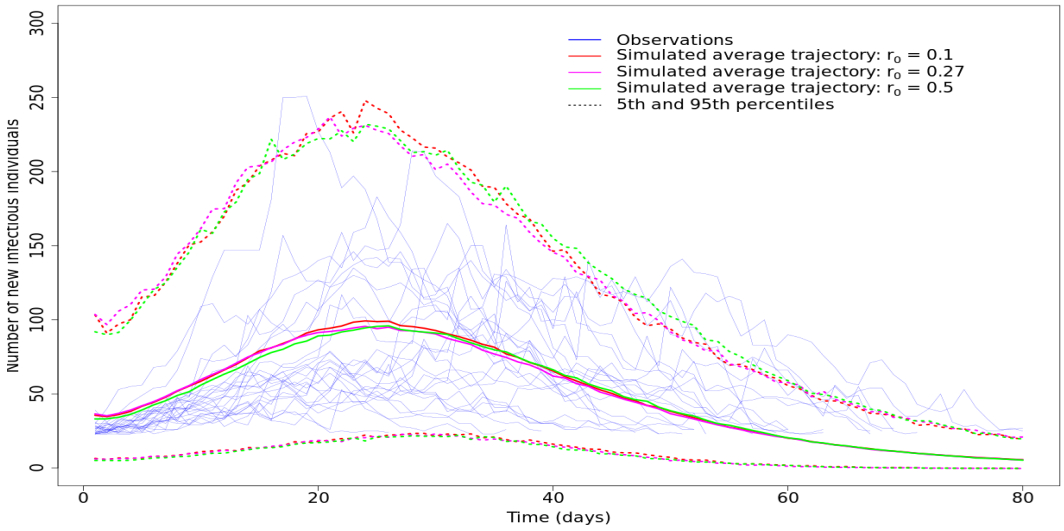

FIGURE 3.5 – Post-predictive check. Observations (number of ILI as proxy for new infectious for each of the U epidemics) (blue). Simulated trajectories obtained for $r_0 = 0.1$ (red), $r_0 = 0.27$ (magenta) and $r_0 = 0.5$ (green) in three steps : (i) generation of 1000 $\hat{\phi}_u$ values based on estimated values of parameters; (ii) given $\hat{\phi}_u$, simulation of 1000 epidemics according to the model (3.19); (iii) computation of average trajectory (solid line) and 5th and 95th percentiles (dotted lines) of the 1000 simulated epidemics. Population size fixed to $N = 100,000$.

The post-predictive check is shown in Figure 3.5. The difference between the average simulated curves obtained with estimated parameter values is negligible according to the r_0 value. Considering the values of $\tilde{R}$, very close in the three scenarios, the proximity of the predicted trajectories is not surprising. Let us emphasize that the majority of the observations are within the predicted envelope (5th and 95th percentiles). Moreover, the predicted average trajectory informs about generic trends of influenza outbreaks : on average, the epidemic peak should be reached around 25 days after the beginning of the outbreak with an incidence of $90/100,000$ inhabitants approximately.

3.6 Discussion

In this article, we propose a generic inference method taking into account simultaneously in a unique model multiple epidemic trajectories and providing estimations of key parameters from incomplete and noisy epidemic data (prevalence or incidence). The framework of the mixed-effects models was used to describe the inter-epidemic variability, whereas the intra-epidemic variability was modeled by an autoregressive Gaussian process. The Gaussian formulation of the epidemic model for prevalence data used in (Narci et al. (2021a)) was extended to the case where incidence data were considered. Then, the SAEM algorithm was coupled with Kalman-like filtering techniques in order to estimate model parameters.

The performances of the estimators were investigated on simulated data of SIR dynamics, under various scenarios, with respect to the parameter values of epidemic and observation processes, the number of epidemics (U), the average number of observations for each of the U epidemics ($\bar{n}$) and the population size (N). The results show that all estimates are close to the true values (reasonable biases), whatever the inter-epidemic variability setting, even for small values of $\bar{n}$ and U. The performances, in term of precision, are improved when increasing U, whereas the bias and

standard deviations of the estimations decrease when increasing $\bar{n}$. We also compared our method with a two-step empirical approach that processes the different data sets separately and combines the individual parameter estimates a posteriori to provide an estimate of inter-epidemic variability (Narci et al. (2021a)). When the number of observations is too low and/or the coefficient of variation of the random effects is high, SAEM-KM clearly outperforms KM.

The proposed inference method was also evaluated on an influenza data set provided by the Réseau Sentinelles, consisting in the daily number of new infectious individuals per 100, 000 inhabitants between 1990 and 2017 in France, using a SEIR compartmental model. Testing different combinations of values for (d_E, d_I) and r_0, we find that $(d_E, d_I) = (1.9, 4.1)$ leads to the best fitting model. Then, irrespective to the r_0 value, we estimated an average value of $\tilde{R} = (1-r_0)R_0$ to be around 1.6. Moreover, we highlighted a non-negligible variability from season to season that is quantitatively assessed. This variability appears especially in the initial conditions (i_0) and the reporting rate (p), as a combined effect of observational uncertainties and differences between seasons. Although to a lesser extent, R_0 also appears to vary between seasons, plausibly reflecting the variability in the transmission rate (λ). Obviously, the estimations can strongly depend on the choice of the compartmental model, the nature and frequency of the observations and the distribution of the random parameters. Our contribution is to propose a finer estimation of the model parameters by taking into account simultaneously all the influenza outbreaks in France for the inference procedure. This leads to an explicit and rigorous estimation of the seasonal variability.

Other methods have been implemented to deal with multiple epidemic dynamics. Bretó et al. (2020) proposed a likelihood-based inference methods for panel data modeled by non-linear partially observed jump processes incorporating unit-specific parameters and shared parameters. Nevertheless, the framework of mixed-effects models was not really investigated. Prague et al. (2020) used an ODE system with mixed effects on the parameters to analyse the first epidemic wave of Covid-19 in various regions in France by inferring key parameters from the daily incidence of infectious ascertained and hospitalized infectious cases. To our knowledge, there are no published studies aiming at the estimation of key parameters simultaneously from several outbreak time series using both a stochastic modeling of epidemic processes and random effects on model parameters.

The main advantage of our method is to propose a direct access to the inter-epidemic variability between multiple outbreaks. Taking into account simultaneously several epidemics in a unique model leads to an improvement of statistical inference compared with empirical methods which consider independently epidemic trajectories. For example, we can mention two experimental settings : (1) the number of epidemics is high but the number of observations per epidemic is low ; (2) the number of observations per epidemic is high but the number of epidemics is low. In such cases, mixed-effects approaches can provide more satisfying estimation results. This benefit more than compensates for the careful calibration of the tuning parameters of the SAEM algorithm.

In some practical cases in epidemiology, it might be difficult to determine whether a parameter is fixed or random. Consequently, our approach could be associated with model selection techniques to inform this choice, using a criterion based on the log-likelihood of observations (see for instance (Delattre et al. (2014)) and (Delattre and Poursat (2020))). This would allow to determine more precisely which parameters reflect inter-individual variability and thus help to better understand the mechanisms underlying this variability. Moreover, we presented a case study on influenza outbreaks, where the variability between epidemics is seasonal, but our approach can be also applied on epidemics spreading simultaneously in many regions. In this case, the inter-epidemic

variability is spatial and it would be interesting to evaluate trends from one region to another.

Chapitre 4

Estimation dans des modèles à effets mixtes : application sur des données épidémiques régionales de la Covid-19 en France

Table des matières

4.1 Introduction **89**
4.1.1 Contexte : pandémie de Covid-19 89
4.1.2 Bref état de l'art des modèles dynamiques Covid-19 en France 89
4.1.3 Objectifs et démarche 90
4.1.4 Description des données 91
4.2 Modélisation compartimentale de propagation de la Covid-19 . . . . . **91**
4.2.1 Modèle mécaniste (formalisme EDO) 92
4.2.2 Expression du R_0 à partir des paramètres du modèle 93
4.3 Inférence **94**
4.3.1 Modèle pour une épidémie 94
4.3.2 Modèles à effets mixtes 97
4.4 Etude empirique de l'identifiabilité du modèle **98**
4.4.1 Design des simulations 98
4.4.2 Simulations hiéararchiques 98
4.4.3 Réglages algorithmiques 99
4.4.4 Résultats 99
4.5 Application à l'épidémie de la Covid-19 dans 12 régions de la France métropolitaine **99**
4.6 Discussion **102**

Note Ce chapitre fera l'objet d'un article.

4.1 Introduction

4.1.1 Contexte : pandémie de Covid-19

En décembre 2019, plusieurs cas de pneumonies sont reportés à Wuhan dans la province de Hubei en Chine. Le 7 janvier 2020, un nouveau coronavirus est identifié, appelé SARS-CoV-2 et responsable de la maladie à coronavirus (Covid-19). Le 24 janvier, trois cas d'infection par la Covid-19 sont détectés en France (Stoecklin et al. (2020)). A partir du 27 février, l'incidence journalière des infections commence à croître exponentiellement, suivie quelques jours plus tard par une augmentation rapide du nombre d'hospitalisations et d'admissions en soins intensifs. Le 11 mars, l'Organisation Mondiale de la Santé déclare une pandémie mondiale de la Covid-19. Dans l'objectif de réduire drastiquement la propagation de la Covid-19, un premier confinement national est annoncé par le gouvernement français du 17 mars au 11 mai.

4.1.2 Bref état de l'art des modèles dynamiques Covid-19 en France

Dans le cas particulier de la France, plusieurs équipes de recherche se sont intéressées à la compréhension de la dynamique épidémique et à son contrôle. Parmi les études (pré)publiées dans la littérature, certaines proposent des approches de modélisation pour fournir des estimations du R_0 et du $R_{\text{eff}}(t)$ (R effectif, *i.e.* le nombre moyen de cas secondaires générés par un individu infectieux à un temps t donné dans une population dont une fraction a déjà été infectée ou est immunisée, e.g. Cazelles et al. (2021)), des prédictions et des études de scénarios (e.g. Liu et al. (2020), Sofonea et al. (2020)) ainsi qu'une évaluation de l'efficacité des mesures de contrôle implémentées (e.g. vaccination : Kiem et al. (2021), Collin et al. (2021), confinement : Prague et al. (2020), Sofonea et al. (2020), Collin et al. (2021), Cazelles et al. (2021), Roux et al. (2021), Di Domenico et al. (2020)) et de l'influence de variants (e.g. variant alpha : Gaymard et al. (2021)) et leur émergence (Blanquart et al. (2021)).

Par exemple, afin d'évaluer l'impact de la mise en place du confinement national instauré le 17 mars 2020 via l'estimation du R_{eff} en fonction du temps, Cazelles et al. (2021) ont utilisé des données d'incidence quotidiennes d'individus infectés, hospitalisés, en réanimation, décédés et sortis de l'hôpital à partir du 28 février 2020 et jusqu'à début 2021 dans cinq régions de la France métropolitaine (Île-de-France, Provence-Alpes-Côte d'Azur, Occitanie, Nouvelle-Aquitaine et Auvergne-Rhône-Alpes). Pour cela, les auteurs considèrent un modèle épidémique SEIR étendu, prenant en compte les individus asymptomatiques (A), hospitalisés (H), en réanimation (ICU) et décédés (D) (cf Figure 6.8 pour une description du modèle à compartiments) défini par des équations différentielles stochastiques et avec de plus un taux de transmission aléatoire $\lambda(t)$ variant dans le temps et modélisé par un processus de diffusion.

Dans une étude rétrospective visant à comparer l'efficacité d'un confinement national sur le nombre d'hospitalisations, de lits de réanimation occupés et de décès par rapport à un confinement ciblé d'un nombre restreint de régions, Roux et al. (2021) ont utilisé un modèle épidémique (cf Figure 6.5) défini par un système d'EDO, structuré en âge et basé sur la démographie et le profil d'âge de la population de 13 régions de la France métropolitaine.

Pour évaluer l'impact du premier confinement national français et l'efficacité de stratégies alternatives de distanciation sociale telles que la fermeture des écoles et l'instauration du télétravail, Di Domenico et al. (2020) ont utilisé un modèle épidémique (cf Figure 6.7) stochastique, structuré en âge et intégrant des données sur le profil d'âge et les contacts sociaux en Île-de-France. En particulier, le ratio de reproduction de base R_0 est estimé à partir des données d'admission à l'hôpital

avant l'instauration du confinement.

En considérant un modèle mécaniste SEIR étendu défini par un système d'EDO incorporant des effets mixtes sur des paramètres (cf Figure 6.6), Prague et al. (2020) ont utilisé les données d'incidence quotidiennes d'infections avérées et d'hospitalisations issues de 12 régions de la France métropolitaine entre le 2 mars et le 11 mai 2020 pour estimer l'impact du confinement sur le taux de transmission.

Enfin, en utilisant un modèle mécaniste similaire (EDO) incluant la couverture vaccinale (cf Figure 6.6), Collin et al. (2021) ont estimé un taux de transmission aléatoire (*i.e.* variable à l'échelle régionale via l'incorporation d'effets mixtes sur le paramètre) et variant dans le temps à partir des données quotidiennes de prévalence et d'incidence d'hospitalisations dans 12 régions de la France métropolitaine entre le 2 mars 2020 et le 28 mars 2021.

4.1.3 Objectifs et démarche

Dans ce chapitre, nous considérons les données publiques d'incidence d'hospitalisation et de décès provenant de la première vague épidémique dans 12 régions en France métropolitaine. Deux objectifs principaux sont ciblés. Un premier objectif, d'ordre méthodologique, est de confronter l'approche d'inférence décrite dans le **Chapitre 3** à un modèle mécaniste plus complexe, comportant plus de variables d'état et de paramètres. Un deuxième objectif, d'ordre applicatif, est d'apporter une évaluation plus précise et explicite de la variabilité entre dynamiques épidémiques à l'échelle des régions via une approche d'inférence mêlant une modélisation stochastique des épidémies et l'ajout d'effets aléatoires sur les paramètres. L'utilisation d'une modélisation stochastique plutôt que déterministe pour les équations d'état est utile afin de mieux prendre en compte les différentes sources de stochasticité démographiques.

Nous reprenons la démarche présentée dans le **Chapitre 3** que nous résumons brièvement. Tout d'abord, pour une région donnée, les dynamiques épidémiques sont décrites par un modèle mécaniste stochastique puis approchées par des processus Gaussiens à petite variance et à temps continu (cf (3.2)). La formulation linéaire définie en (3.14) pour des données d'incidence est considérée pour la modélisation des variables d'états. Ensuite, un modèle d'observation et son approximation Gaussienne s'écrivant sous la forme (3.16) sont proposés afin de prendre en compte les erreurs de reports dans les données de la Covid-19. Nous combinons ces deux approximations Gaussiennes pour décrire le modèle statistique via une représentation hiérarchique à deux niveaux. Le premier niveau définit le modèle à espace d'état Gaussien et linéaire de la forme (3.19) et permet de décrire la variabilité intra-épidémie. Le deuxième niveau comporte une description de la variabilité inter-épidémies en spécifiant des distributions pour les paramètres du modèle (cf (3.20)). Enfin, l'approche d'inférence combinant l'algorithme SAEM avec des techniques de filtrage de Kalman est utilisée pour estimer les paramètres impliqués dans ces distributions.

La Section 4.2 présente le modèle mécaniste utilisé pour décrire les dynamiques épidémiques. Dans la Section 4.3 est décrit le modèle statistique utilisé pour l'inférence via le cadre des modèles à effets mixtes. La Section 4.4 propose une étude empirique de la méthode d'inférence développée dans le **Chapitre 3** sur des données simulées sous le modèle envisagé dans le cas de la Covid-19. Enfin, les résultats d'estimation sur les données régionales de la Covid-19 en France sont présentés dans la Section 4.5.

4.1.4 Description des données

Durant la première vague, les nombres quotidiens de nouvelles infections dues à la Covid-19 étaient fortement sous-estimés en raison du faible nombre de tests disponibles et de la présence d'individus asymptomatiques (et donc non-reportés, Pullano et al. (2021)). Les données d'incidence hospitalières, comprenant en particulier les nouvelles hospitalisations et les nouveaux décès, semblent alors plus fiables et moins sujettes à des erreurs de report. Nous utilisons les données correspondant aux nombres quotidiens d'individus hospitalisés et décédés dans les 12 régions métropolitaines de France, disponibles à partir du 19 mars 2020 (cf Figure 4.1). Celles-ci sont disponibles sur une plateforme ouverte des données publiques françaises :
(url : **www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/**).
Comme les données hospitalières sont disponibles seulement après la mise en place du premier confinement, nous considérons aussi les reports quotidiens de nouvelles infections entre le 28 février et le 18 mars 2020 pour chaque région. Ces données sont disponibles au lien suivant :
(url : **www.data.gouv.fr/fr/datasets/chiffres-cles-concernant-lepidemie-de-covid19-en-france/**).

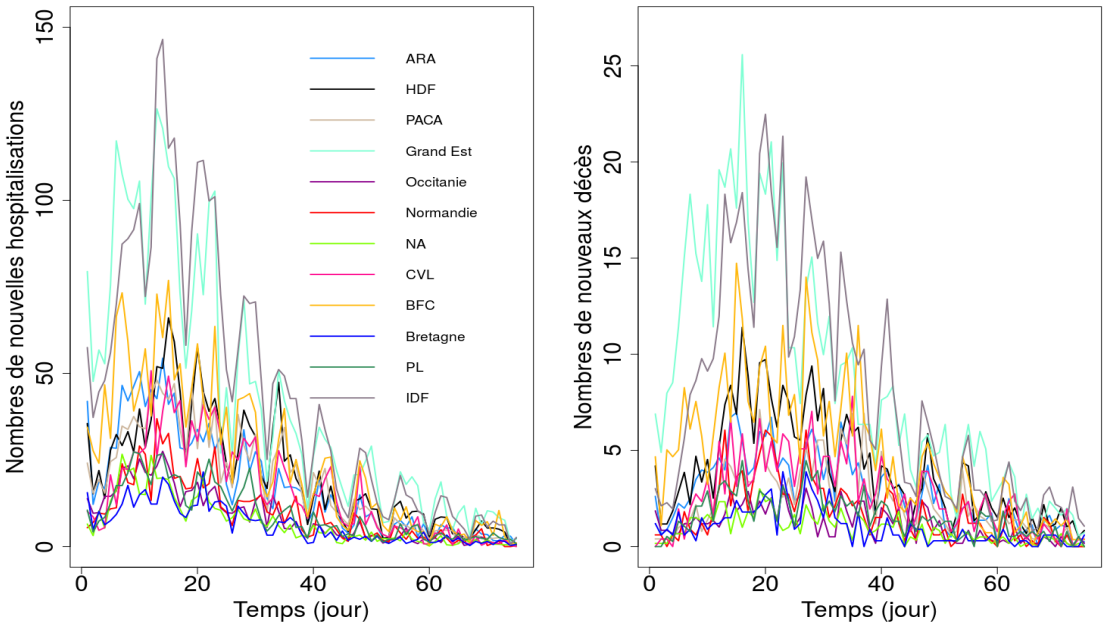

FIGURE 4.1 – Incidence quotidienne d'individus d'hospitalisés (gauche) et décédés (droite), calculée sur $1,000,000$ d'habitants, dans 12 régions (une couleur par région) de la France métropolitaine entre le 19 mars et le 1er juin 2020. ARA = Auvergne-Rhône-Alpes, HDF = Hauts-de-France, PACA = Provence-Alpes-Côte d'Azur, NA = Nouvelle-Aquitaine, CVL = Centre-Val de Loire, BFC = Bourgogne-Franche-Comté, PL = Pays de la Loire, IDF = Île-de-France.

4.2 Modélisation compartimentale de propagation de la Covid-19

Une première réflexion concerne le choix du modèle mécaniste pour décrire les dynamiques épidémiques. En particulier, un équilibre doit être trouvé entre la complexité du modèle d'une part (*i.e.* le nombre de paramètres impliqués) et sa capacité à décrire fidèlement les différentes dynamiques d'autre part. Pour cela, nous nous inspirons du cadre étendu du modèle SEIR proposé par Pan et al. (2020), prenant en compte la présence d'individus asymptomatiques et d'individus

hospitalisés, auquel nous ajoutons un compartiment correspondant aux décès. Nous parlons alors du modèle SEIRAHD.

4.2.1 Modèle mécaniste (formalisme EDO)

Le modèle à compartiments SEIRAHD partitionne la population de taille N en individus susceptibles (S), exposés (E), infectieux et symptomatiques (I), infectieux et asymptomatiques (A), hospitalisés (H), décédés (D) et guéris (R). Dans la suite, par convention, nous utilisons la lettre majuscule S (resp. E, I, etc.) pour désigner l'effectif d'individus susceptibles (resp. exposés, infectieux, etc.) et la lettre minuscule s (resp. e, i, etc.) pour désigner la proportion d'individus susceptibles (resp. exposés, infectieux, etc.).

Dans ce modèle, nous supposons que les individus hospitalisés ne peuvent plus transmettre la maladie et que le mélange entre les individus est homogène dans la population (hypothèse plausible à l'échelle régionale). La transmission par les individus infectieux symptomatiques et asymptomatiques est régie par un taux de transmission noté λ. Une fois infecté et après une durée d'exposition $d_E = 1/\epsilon$, l'individu a une probabilité τ_A d'être asymptomatique et $1 - \tau_A$ d'être symptomatique. De plus, nous considérons que la force de transmission d'un individu asymptomatique est réduit d'un facteur $0 < \alpha < 1$ par rapport à celle d'un individu symptomatique. Un individu symptomatique est hospitalisé après une durée $d_I = 1/\gamma$ suite à l'infection avec probabilité τ_H ou guérit avec probabilité $1 - \tau_H$. Nous supposons aussi que les individus infectieux symptomatiques et infectieux asymptomatiques guérissent après une durée d_I. Enfin, après une durée d'hospitalisation moyenne $d_H = 1/\kappa$, un individu hospitalisé meurt avec probabilité τ_D ou guérit avec probabilité $1 - \tau_D$. Le système d'EDO (4.1) ci-dessous formalise les dynamiques des proportions d'individus dans chaque compartiment (voir Figure 4.2 pour un schéma). Dans la suite, par souci de lisibilité, la dépendance en temps est omise lorsqu'il n'y a pas d'ambiguïté. Bien que le modèle mécaniste soit décrit dans le formalisme déterministe, une description stochastique de celui-ci est utilisée dans la suite.

$$\begin{cases} \frac{ds}{dt} & = -\lambda s(i + \alpha a), \\ \frac{de}{dt} & = \lambda s(i + \alpha a) - \epsilon e, \\ \frac{di}{dt} & = (1 - \tau_A)\epsilon e - \gamma i, \\ \frac{da}{dt} & = \tau_A \epsilon e - \gamma a, \\ \frac{dh}{dt} & = \tau_H \gamma i - \kappa h, \\ \frac{dd}{dt} & = \tau_D \kappa h, \\ r & = 1 - s - e - i - a - h - d, \\ x_0 & = (s_0, e_0, i_0, r_0, a_0, h_0, d_0)^t \neq (0, \ldots, 0)^t. \end{cases} \tag{4.1}$$

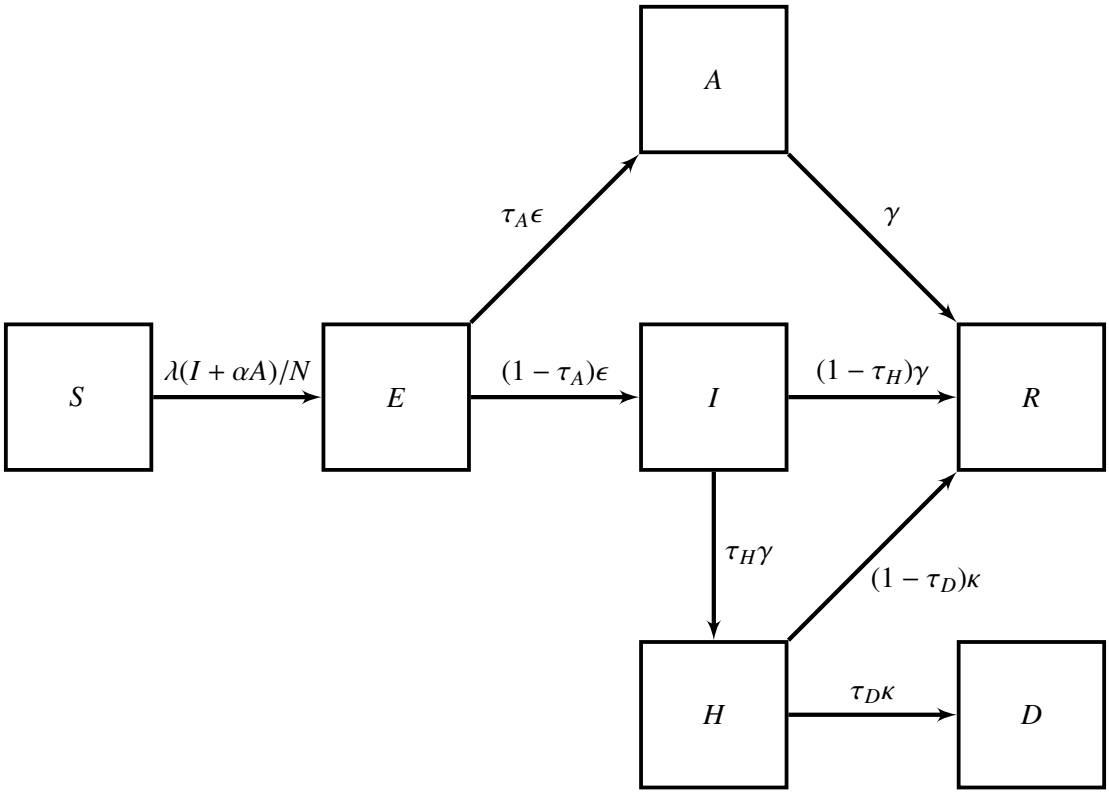

FIGURE 4.2 – Modèle à compartiments SEIRAHD avec sept blocs représentant respectivement les effectifs au sein d'une population de taille N d'individus susceptibles (S), exposés (E), infectieux symptomatiques (I), infectieux asymptomatiques (A), hospitalisés (H), décédés (D) et guéris (R). Les transitions des individus d'un état de santé à un autre sont gouvernées par le taux de transmission λ, le taux d'exposition ϵ, le taux de guérison γ, la durée d'hospitalisation $1/\kappa$, la fraction d'individus asymptomatiques τ_A, hospitalisés τ_H et décédés τ_D et le facteur de réduction de transmission chez les individus asymptomatiques α.

En notant $d_E = \frac{1}{\epsilon}$ (resp. $d_I = \frac{1}{\gamma}$ et $d_H = \frac{1}{\kappa}$) la durée d'exposition (resp. la durée d'infectiosité et d'hospitalisation), le vecteur des paramètres du modèle mécaniste, incluant la condition initiale x_0, correspond à :

$$\eta = (\lambda, d_E, d_I, d_H, \tau_A, \tau_H, \tau_D, \alpha, x_0).$$

4.2.2 Expression du R_0 à partir des paramètres du modèle

Un indicateur épidémique d'intérêt, le ratio de reproduction de base R_0, peut s'exprimer en fonction de ces paramètres en calculant la matrice de nouvelle génération du modèle (e.g. van den Driessche and Watmough (2002)). Nous détaillons ci-dessous les différentes étapes dans le cas du modèle SEIRAHD. Tout d'abord, il faut déterminer le nombre d'états d'infection. Dans le modèle (4.1), il y en a 3 : $z = (e, i, a)$ avec $z_1 = e$, $z_2 = i$ et $z_3 = a$. Nous cherchons alors à écrire $\frac{dz_j}{dt} = \mathcal{F}_j(z) + v_j^+(z) - v_j^-(z)$ pour $j \in \{1, 2, 3\}$ où $\mathcal{F}_j(z)$, $v_j^+(z)$ et $v_j^-(z)$ désignent respectivement le taux d'apparition des nouvelles infections dans le compartiment z_j, le taux de transfert des individus entrant dans le compartiment z_j pour toute autre raison que l'infection et le taux de transfert des individus sortant du compartiment z_j :

$$\begin{cases} \mathcal{F}_1(z) = \lambda s(i + \alpha a), \quad \mathcal{F}_2(z) = 0, \quad \mathcal{F}_3(z) = 0, \\ v_1^+(z) = 0, \quad v_2^+(z) = (1 - \tau_A)\epsilon e, \quad v_3^+(z) = \tau_A \epsilon e, \\ v_1^-(z) = \epsilon e, \quad v_2^-(z) = \gamma i, \quad v_3^-(z) = \gamma a. \end{cases}$$

Pour $1 \leq j, l \leq 3$, soient :

$$F = \frac{\partial \mathcal{F}_j(z)}{\partial z_l} = \begin{pmatrix} 0 & \lambda s & \lambda \alpha s \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

et

$$V = \frac{\partial \left(v_j^-(z) - v_j^+(z) \right)}{\partial z_l} = \begin{pmatrix} \epsilon & 0 & 0 \\ -(1-\tau_A)\epsilon & \gamma & 0 \\ -\tau_A \epsilon & 0 & \gamma \end{pmatrix}.$$

Alors, R_0 est égal à la plus grande valeur propre (en valeur absolue) de la matrice FV^{-1} en $z = z^*$, où z^* est l'état d'équilibre sans maladie, *i.e.* $s^* = 1$ et $i^* = e^* = a^* = h^* = d^* = r^* = 0$. Nous avons :

$$FV^{-1} = \begin{pmatrix} \frac{\lambda s^*(1-\tau_A+\tau_A\alpha)}{\gamma} & \frac{\lambda s^*}{\gamma} & \frac{\lambda \alpha s^*}{\gamma} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Les valeurs propres d'une matrice triangulaire supérieure correspondant aux éléments de sa diagonale et sachant $s^* = 1$, nous en déduisons :

$$R_0 = \frac{\lambda}{\gamma}(1 - \tau_A + \tau_A \alpha) = R_0^A + R_0^I, \tag{4.2}$$

avec $R_0^A = \frac{\lambda}{\gamma}(1 - \tau_A)$ et $R_0^I = \frac{\lambda}{\gamma}\tau_A \alpha$.

Une autre manière de retrouver l'expression 4.2 consiste à partir de la définition du R_0, correspondant au nombre moyen de cas secondaires générés par un individu infectieux dans une population entièrement susceptible. Dans le cas du modèle SEIRAHD, le R_0 s'écrit comme la somme du R_0^A due aux infections par les individus asymptomatiques et R_0^I due aux infections par les individus symptomatiques. Sachant qu'un individu infecté a une probabilité $(1 - \tau_A)$ d'être symptomatique, il peut transmettre la maladie au taux λ à un nombre moyen d'individus susceptibles égal à $R_0^I = \frac{\lambda}{\gamma}(1-\tau_A)$. De plus, pour une probabilité τ_A, un individu infectieux est asymptomatique et peut transmettre la maladie au taux $\alpha\lambda$ à un nombre moyen d'individus susceptibles égal à $R_0^A = \frac{\lambda}{\gamma}\alpha\tau_A$. En additionnant ces deux nombres, nous retrouvons bien l'expression (4.2).

4.3 Inférence

Nous reprenons la même démarche que celle présentée dans le **Chapitre 3** visant à proposer une modélisation hiérarchique des dynamiques épidémiques via la description de la variabilité intra-épidémie et inter-épidémies. Pour rappel, la première variabilité est décrite par le modèle stochastique approché à espace d'état Gaussien et linéaire (3.19), issu d'une approximation Gaussienne du modèle de processus de sauts et du modèle des observations. La deuxième variabilité est quant à elle prise en compte via la spécification d'une distribution pour les paramètres (aléatoires) du modèle (cf 3.20).

4.3.1 Modèle pour une épidémie

Dans la suite, chaque épidémie $X_u(t) = \left(\frac{S_u(t)}{N_u}, \frac{E_u(t)}{N_u}, \frac{I_u(t)}{N_u}, \frac{A_u(t)}{N_u}, \frac{H_u(t)}{N_u}, \frac{D_u(t)}{N_u} \right)^t$, $u = 1, \ldots, U$ avec $U = 12$ régions et N_u la taille de la population de la région u, est observée à des temps discrets fixes et régulièrement espacés $t_0 = 0 < t_1 < \ldots < t_{n_u}$, avec n_u le nombre d'observations pour l'épidémie u et $t_k = k\Delta$ où $\Delta = 1$ est le pas de temps quotidien. De plus, nous notons

$$\eta_u = (\lambda_u, d_{E,u}, d_{I,u}, d_{H,u}, \tau_{A,u}, \tau_{H,u}, \tau_{D,u}, \alpha_u, x_{0,u}), \text{ avec } x_{u,0} = \frac{X_u(0)}{N_u},$$

le vecteur des paramètres du modèle spécifiques à l'épidémie u.

4.3.1.1 Modélisation statistique des variables d'état

Soit $X_{u,k} := X_u(t_k)$ pour $k \geq 0$. Comme les observations disponibles correspondent à des données d'incidence, les variables d'état sont assimilées aux accroissements de $(X_{u,k})_{k\geq 1}$, notés $(\Delta_k X_u)_{k\geq 1}$, et exprimées comme suit :

$$\Delta_k X_u := X_{u,k} - X_{u,k-1} = G_k(\eta_u) + (A_{k-1}(\eta_u) - I_6) \sum_{l=1}^{k-1} \Delta_l X_u + V_{u,k}, \tag{4.3}$$

avec $V_{u,k} \sim \mathcal{N}_6\,(0, T_k(\eta_u, \Delta))$ et $G_k(\cdot)$, $A_{k-1}(\cdot)$ et $T_k(\cdot)$, définis en (3.15), (3.6) et (3.8), sont obtenus à partir de la fonction de dérive $b(\cdot)$ (et de son gradient) et de la matrice de diffusion $\Sigma(\cdot)$ définis en (2.2).

Ici, pour tout $x \in [0,1]^6$, la fonction de dérive $b(\eta, \cdot)$ et la matrice de diffusion $\Sigma(\eta, \cdot)$ s'écrivent respectivement :

$$b(\eta, x) = \begin{pmatrix} -\lambda s(i + \alpha a) \\ \lambda s(i + \alpha a) - \epsilon e \\ (1 - \tau_A)\epsilon e - \gamma i \\ \tau_A \epsilon e - \gamma a \\ \tau_H \gamma i - \kappa h \\ \tau_D \kappa h \end{pmatrix};$$

$$\Sigma(\eta, x) = \begin{pmatrix} \lambda s(i + \alpha a) & -\lambda s(i + \alpha a) & 0 & 0 & 0 & 0 \\ -\lambda s(i + \alpha a) & \lambda s(i + \alpha a) + \epsilon e & -(1 - \tau_A)\epsilon e & -\tau_A \epsilon e & 0 & 0 \\ 0 & -(1 - \tau_A)\epsilon e & (1 - \tau_A)\epsilon e + \gamma i & 0 & -\tau_H \gamma i & 0 \\ 0 & -\tau_A \epsilon e & 0 & \tau_A \epsilon e + \gamma a & 0 & 0 \\ 0 & 0 & -\tau_H \gamma i & 0 & \tau_H \gamma i + \tau_D \kappa h & -\tau_D \kappa h \\ 0 & 0 & 0 & 0 & -\tau_D \kappa h & \tau_D \kappa h \end{pmatrix}.$$

4.3.1.2 Modélisation statistique des observations

Nous considérons que le temps d'observation $t_0 = 0$ correspond au 27 février. Pour $t_k = 1, \ldots, 20$, soit du 28 février au 18 mars, nous utilisons les reports quotidiens de nouvelles infections. Pour $t_k = 21, \ldots, 95$, soit du 19 mars au 1er juin, les données utilisées sont celles correspondant aux nombres de nouveaux hospitalisés et de nouveaux décès, cela faisant 75 observations quotidiennes.

Dans le modèle SEIRAHD, la proportion de nouvelles infections $O_{I,u}(t_k)$ aux temps t_k, $k = 1, \ldots, 20$, dans la région u peut être exprimée par :

$$\begin{aligned} O_{I,u}(t_k) &= \int_{t_{k-1}}^{t_k} (1 - \tau_A)\epsilon e_u(t)\, dt \\ &= -(1 - \tau_A)(\Delta_k s_u + \Delta_k e_u). \end{aligned}$$

Les proportions de nouvelles hospitalisations $O_{H,u}(t_k)$ et de nouveaux décès $O_{D,u}(t_k)$ aux temps t_k,

$k = 21, \ldots, 95$, dans la région u peuvent être respectivement assimilées à

$$\begin{aligned} O_{H,u}(t_k) &= \int_{t_{k-1}}^{t_k} \tau_H \gamma i_u(t)\, dt \\ &= \int_{t_{k-1}}^{t_k} \left[\frac{dh_u}{dt} + \kappa h_u(t) \right] dt \\ &= \int_{t_{k-1}}^{t_k} \left[\frac{dh_u}{dt} + \frac{1}{\tau_D} \frac{dd_u}{dt} \right] dt \\ &= \Delta_k h_u + \frac{1}{\tau_D} \Delta_k d_u, \end{aligned}$$

et

$$O_{D,u}(t_k) = \int_{t_{k-1}}^{t_k} \tau_D \kappa h_u(t)\, dt = \Delta_k d_u.$$

Notons $(Y_{u,k})_{k \geq 1}$ des observations bruitées de $O_{I,u}(t_k)$, $O_{H,u}(t_k)$ et $O_{D,u}(t_k)$. Pour modéliser les observations, nous considérons :

$$\begin{cases} Y_{u,k} & = Y_{u,k}^{(1)} \mathbf{1}_{1 \leq k \leq 20} + Y_{u,k}^{(2)} \mathbf{1}_{k > 20}, \\ Y_{u,k}^{(1)} & \sim \mathcal{N}\left(\rho_{I,u} O_{I,u}(t_k), \frac{1}{N_u} \rho_{I,u} O_{I,u}(t_k) \right), \\ Y_{u,k}^{(2)} & \sim \mathcal{N}_2\left(\begin{pmatrix} O_{H,u}(t_k) \\ O_{D,u}(t_k) \end{pmatrix}, \frac{1}{N_u} \begin{pmatrix} O_{H,u}(t_k) & 0 \\ 0 & O_{D,u}(t_k) \end{pmatrix} \right), \end{cases} \tag{4.4}$$

où $\rho_{I,u}$ est le taux de reports d'individus nouvellement infectieux dans la région u.

4.3.1.3 Stratégie d'estimation des paramètres du modèle

Comme les modèles des états (4.3) et des observations (4.4) contiennent beaucoup de paramètres (16 en tout, cf Tableau 4.1), il s'agit de trouver un compromis entre le nombre de paramètres dont il est vraisemblable de considérer la valeur connue (e.g. valeur mesurée expérimentalement avec une grande fiabilité, valeur estimée par d'autres études à partir d'autres données) et le nombre de paramètres qu'il est possible d'estimer (e.g. par une étude sur données simulées) à partir des données disponibles. La stratégie adoptée est présentée dans le Tableau 4.1, en nous inspirant de plusieurs études (cf Tableau 6.8 pour un résumé succint de ces études).

Dans le cas de la fraction d'asymptomatiques τ_A et de la durée d'hospitalisation d_H, comme les valeurs figurant dans la littérature sont constrastées (cf Tableau 6.8) et qu'ils sont difficilement estimables, nous testons trois valeurs pour chacun de ces paramètres : $\tau_A \in \{0.2, 0.4, 0.6\}$ et $d_H \in \{10, 15, 20\}$ jours. Pour une région u, nous exprimons la proportion initiale d'individus exposés $e_{0,u}$ comme la solution de l'équation pour la dynamique de la variable e_u à l'équilibre :

$$\left. \frac{de_u}{dt} \right|_{t=0} = \lambda_u s_{0,u} (i_{0,u} + \alpha a_{0,u}) - \frac{e_{0,u}}{d_E} = 0,$$

soit

$$e_{0,u} = \lambda_u d_E (i_{0,u} + \alpha a_{0,u}),$$

sous l'hypothèse $s_{0,u} \approx 1$. A partir de (4.1), la proportion initiale d'individus asymptomatiques $a_{0,u}$ peut s'exprimer en fonction de $i_{0,u}$ et τ_A :

$$a_{0,u} = \frac{\tau_A}{1 - \tau_A} i_{0,u}.$$

En faisant l'hypothèse que $h_{0,u} = d_{0,u} = r_{0,u} = 0$ pour tout u, il en résulte que connaître $i_{0,u}$ revient à connaître $e_{0,u}$ et $a_{0,u}$.

TABLE 4.1 – Paramètres du modèle SEIRAHD (4.1) et du modèle des observations (4.4). Les valeurs fixées de certains paramètres sont déduites du contexte épidémiologique.

Paramètre	Signification	Valeur	Référence
λ	Taux de transmission	Aléatoire	**Estimé**
d_E	Durée d'exposition (jours)	4	Cazelles et al. (2021)
d_I	Durée d'infectiosité (jours)	6	Cazelles et al. (2021)
d_H	Durée d'hospitalisation (jours)	$\{10, 15, 20\}$	cf Tableau 6.8
τ_A	Fraction d'individus asymptomatiques	$\{0.2, 0.4, 0.6\}$	cf Tableau 6.8
τ_H	Fraction d'individus hospitalisés	Aléatoire	**Estimé**
τ_D	Fraction d'individus décédés	Aléatoire	**Estimé**
α	Réduction de transmission	0.55	Li et al. (2020)
s_0	Proportion initiale d'individus susceptibles	≈ 1	Fixé
e_0	Proportion initiale d'individus exposés	Aléatoire	**Estimé**
i_0	Proportion initiale d'infectieux symptomatiques	Aléatoire	**Estimé**
r_0	Proportion initiale d'individus guéris	0	Fixé
a_0	Proportion initiale d'infectieux asymptomatiques	Aléatoire	**Estimé**
h_0	Proportion initiale d'individus hospitalisés	0	Fixé
d_0	Proportion initiale d'individus décédés	0	Fixé
ρ_I	Taux de reports de nouvelles infections	Aléatoire	**Estimé**

4.3.2 Modèles à effets mixtes

Nous proposons une modélisation hiérarchique à deux niveaux pour prendre en compte la variabilité intra-épidémie et inter-épidémies. Dans la suite, nous notons $\phi_u = (\eta_u, \rho_{I,u})$ le vecteur des paramètres spécifiques à l'épidémie de la région u.

4.3.2.1 Variabilité intra-épidémie

Conditionnellement à $\phi_u = \varphi$, nous considérons le modèle à espace d'état ci-dessous, comprenant une modélisation des variables d'états (cf (4.3)) et une modélisation des observations (cf (4.4)) :

$$\begin{cases} \Delta_k X_u = G_k(\varphi) + (A_{k-1}(\varphi) - I_d) \sum_{l=1}^{k-1} \Delta_l X_u + V_{u,k}, \quad k \geq 1, \\ Y_{u,k} = \left[\tilde{B}^{(1)}(\varphi)\mathbf{1}_{1\leq k\leq 20} + \tilde{B}^{(2)}(\varphi)\mathbf{1}_{k>20}\right] \Delta_k X_u + \tilde{W}^{(1)}_{u,k}\mathbf{1}_{1\leq k\leq 20} + \tilde{W}^{(2)}_{u,k}\mathbf{1}_{k>20}, \end{cases} \tag{4.5}$$

où $\mathbf{1}$ est la fonction indicatrice, $\tilde{W}^{(1)}_{u,k} \sim \mathcal{N}\left(0, \tilde{P}^{(1)}_k(\varphi)\right)$, $\tilde{W}^{(2)}_{u,k} \sim \mathcal{N}_2\left(0, \tilde{P}^{(2)}_k(\varphi)\right)$ tels que

$$\begin{aligned} \tilde{B}^{(1)}(\phi_u) &= \begin{pmatrix} -(1-\tau_A)\rho_I & -(1-\tau_A)\rho_I & 0 & 0 & 0 & 0 \end{pmatrix}, \\ \tilde{B}^{(2)}(\phi_u) &= \begin{pmatrix} \tilde{B}_H(\phi_u) \\ \tilde{B}_D(\phi_u) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & \frac{1}{\tau_{D,u}} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \\ \tilde{P}^{(1)}_k(\phi_u) &= \frac{1}{N_u}\tilde{B}^{(1)}(\phi_u)\Delta_k x_u, \\ \tilde{P}^{(2)}_k(\phi_u) &= \begin{pmatrix} \frac{1}{N_u}\tilde{B}_H(\phi_u)\Delta_k x_u & 0 \\ 0 & \frac{1}{N_u}\tilde{B}_D(\phi_u)\Delta_k x_u \end{pmatrix}, \end{aligned}$$

avec $\Delta_k x_u = x(\eta_u, t_k) - x(\eta_u, t_{k-1})$ et $x(\cdot, t)$ la solution déterministe du système (4.1).

4.3.2.2 Variabilité inter-épidémies

Nous supposons que les paramètres spécifiques à chaque épidémie (ϕ_u) sont des variables aléatoires indépendantes et identiquement distribuées selon

$$\begin{cases} \phi_u & = (\lambda_u, \tau_{H,u}, \tau_{D,u}, i_{0,u}, \rho_{I,u})^t = h(\beta, \xi_u), \quad \text{avec} \\ h_1(\beta, \xi_u) & = \exp\left[\beta_1 + \xi_{1,u}\right], \\ h_j(\beta, \xi_u) & = \frac{1}{1+\exp\left[-(\beta_j + \xi_{j,u})\right]}, \ j = 2, 3, 4, 5, \end{cases} \tag{4.6}$$

où $\beta \in \mathbb{R}^5$ est un vecteur d'effets fixes et $\xi_u \sim_{i.i.d.} \mathcal{N}_5(0, \Gamma)$ sont des effets aléatoires avec Γ une matrice de covariance supposée diagonale.

4.4 Etude empirique de l'identifiabilité du modèle

Dans cette section, nous présentons les résultats numériques obtenus sur des données simulées par la méthode d'inférence combinant l'algorithme SAEM et du filtrage de Kalman (SAEM-KM). En particulier, l'étude des performances de la méthode sur des données simulées permet d'évaluer quels sont les paramètres identifiables du modèle sachant les observations disponibles.

4.4.1 Design des simulations

Les valeurs des paramètres connus que nous utilisons sont celles présentées dans le Tableau 4.1. De plus, nous posons $\tau_A = 0.2$ et $d_H = 20$ jours. Nous considérons les valeurs suivantes pour les effets fixes β et les éléments de la matrice de covariance Γ :

$$\begin{aligned} \beta &= (-0.23, -4.60, -2.20, -7.26, -2.94)', \\ \Gamma &= \text{diag}(0.15^2, 0.5^2, 0.5^2, 0.5^2, 0.5^2), \end{aligned}$$

ce qui implique :

$$\begin{aligned} \mathbb{E}(\lambda_u) &\approx 0.800, \quad CV_{\lambda_u} \approx 15\%; \\ \mathbb{E}(\tau_{H,u}) &\approx 0.011, \quad CV_{\tau_{H,u}} \approx 52\%; \\ \mathbb{E}(\tau_{D,u}) &\approx 0.109, \quad CV_{\tau_{D,u}} \approx 45\%; \\ \mathbb{E}(i_{0,u}) &\approx 7.9.10^{-4}, \quad CV_{i_{0,u}} \approx 53\%; \\ \mathbb{E}(\rho_{I,u}) &\approx 0.055, \quad CV_{\rho_{I,u}} \approx 49\%; \end{aligned}$$

où CV_ϕ correspond au coefficient de variation de la variable aléatoire ϕ.

4.4.2 Simulations hiéararchiques

Nous considérons une taille de population $N_u = N = 5,000,000$ fixe. Pour $U = 12$, $J = 100$ jeux de données sont générés, chacun étant composé de U trajectoires épidémiques décrites par le modèle SEIRAHD. Dans un premier temps, des échantillons indépendants de

$$\left(\phi_{u,j} = (\lambda_u, \tau_{H,u}, \tau_{D,u}, i_{0,u}, \rho_{I,u})'_j\right), u = 1, \dots U, j = 1, \dots, J,$$

sont tirés selon le modèle (4.6). Dans un deuxième temps, conditionnellement à chaque jeu de paramètres $\phi_{u,j}$, un processus Markovien de sauts de dimension 6

$$\mathcal{Z}_{u,j}(t) = (S_{u,j}(t), E_{u,j}(t), I_{u,j}(t), A_{u,j}(t), H_{u,j}(t), D_{u,j}(t))',$$

correspondant à la version stochastique du système d'EDO (4.1) sur les effectifs, est simulé à partir de l'algorithme de Gillespie (Gillespie (1977)). Normaliser le processus $\mathcal{Z}_{u,j}(t)$ par N et le discrétiser en des temps réguliers $t_k = k\Delta$, $k = 1, \ldots, n_{u,j}$, permet d'obtenir les trajectoires $X_{u,k,j} = \left(\frac{S_{u,k,j}}{N}, \frac{E_{u,k,j}}{N}, \frac{I_{u,k,j}}{N}, \frac{A_{u,k,j}}{N}, \frac{H_{u,k,j}}{N}, \frac{D_{u,k,j}}{N}\right)^t$. Un pas de temps fixe $\Delta = 1$ et un nombre d'observations fixe $n_{u,j} = 75$ pour tout u et j sont considérés. Enfin, conditionnellement aux trajectoires simulées $X_{u,k,j}$ et aux paramètres $\phi_{u,j}$, les observations $Y_{u,k,j}$ sont générées selon le modèle d'observation (4.4).

4.4.3 Réglages algorithmiques

Pour des considérations d'implémentation, nous référons le lecteur au **Chapitre 6**, Section B.3. En particulier, pour l'algorithme SAEM, les paramètres de réglage algorithmiques ont été choisis comme suit : le nombre d'itérations de chauffe $M_0 = 1000$, $\nu_0 = 0.6$ le paramètre apparaissant dans l'expression du pas de discrétisation à l'étape m de l'algorithme SAEM $\alpha_m = \frac{1}{(m-M_0)^{\nu_0}}$ avec $m > M_0$, $K_0 = 0.87$, le seuil de tolérance de convergence des paramètres estimés $\mu_0 = 0.0001$ et le paramètre impliqué dans la méthode du recuit simulé $\tau_0 = 0.98$. L'algorithme s'arrête une fois la convergence atteinte.

4.4.4 Résultats

Le Tableau 4.2 représente les estimations de l'espérance et de l'écart-type des effets mixtes ϕ_u, calculées à partir des estimations de β et Γ en utilisant les fonctions h définies en (4.6). Pour chaque paramètre, les valeurs reportées correspondent à la moyenne des $J = 100$ estimations de paramètres $\phi_{u,j}$, $j \in \{1, \ldots, J\}$, et leurs écarts-type entre parenthèses.

TABLE 4.2 – Valeurs estimées des paramètres obtenues par la méthode d'inférence SAEM-KM. Pour chaque paramètre du modèle, les estimations ponctuelles sont calculées comme la moyenne des $J = 100$ estimations individuelles (écarts-type entre parenthèses).

Paramètres	$\mathbb{E}(\lambda_u)$	$\mathbb{E}(\tau_{H,u})$	$\mathbb{E}(\tau_{D,u})$	$\mathbb{E}(i_{0,u})$	$\mathbb{E}(\rho_{I,u})$	$\mathrm{sd}(\lambda_u)$	$\mathrm{sd}(\tau_{H,u})$	$\mathrm{sd}(\tau_{D,u})$	$\mathrm{sd}(i_{0,u})$	$\mathrm{sd}(\rho_{I,u})$
Vraies valeurs	**0.800**	**0.011**	**0.109**	**0.00079**	**0.055**	**0.120**	**0.006**	**0.049**	**0.00042**	**0.027**
Estimation										
Moyenne	0.780	0.012	0.113	0.00124	0.057	0.113	0.006	0.048	0.00071	0.029
Écart-type	*(0.033)*	*(0.002)*	*(0.015)*	*(0.00024)*	*(0.010)*	*(0.041)*	*(0.003)*	*(0.011)*	*(0.00043)*	*(0.013)*

Le Tableau 4.2 montre que le biais d'estimation pour chaque paramètre est faible et que la précision des estimations est tout à fait satisfaisante. Il semble donc raisonnable de tenter d'estimer les paramètres de ce modèle sur les données de la Covid-19.

4.5 Application à l'épidémie de la Covid-19 dans 12 régions de la France métropolitaine

Nous appliquons maintenant la méthode SAEM-KM aux données réelles de la Covid-19 présentées en Section 4.1.4 en utilisant les modèles (4.5)-(4.6). Nous considérons les mêmes paramètres de réglage algorithmique que ceux évoqués dans la Section 4.4, mais avec $M_0 = 5000$. Nous testons trois valeurs pour $d_H \in \{10, 15, 20\}$ jours et $\tau_A \in \{0.2, 0.4, 0.6\}$, ce qui revient à considérer neuf modèles en tout. En utilisant des techniques d'échantillonnage préférentiel, la log-vraisemblance des observations est estimée stochastiquement pour chaque modèle à partir des

valeurs estimées des paramètres initialement obtenues avec la méthode SAEM-KM (cf Tableau 4.3).

TABLE 4.3 – Valeurs estimées de la log-vraisemblance des observations en testant trois valeurs pour $d_H \in \{10, 15, 20\}$ jours et $\tau_A \in \{0.2, 0.4, 0.6\}$.

d_H (jours)	τ_A	log-vraisemblance estimée
10	0.2	19144.41
	0.4	19136.11
	0.6	19134.19
15	0.2	19607.50
	0.4	19601.41
	0.6	19581.35
20	0.2	19765.34
	0.4	19739.76
	0.6	19708.76

Quelle que soit la valeur pour d_H, les valeurs des log-vraisemblances estimées en fonction de τ_A sont très proches. La différence est plus marquée entre les trois valeurs de log-vraisemblance correspondant aux trois valeurs de d_H à τ_A fixé, avec un maximum de log-vraisemblance estimée atteint en $d_H = 20$ jours.

Dans le cas où $d_H = 20$ jours, le Tableau 4.4 présente les estimations des paramètres du modèle obtenues par la méthode SAEM-KM selon les valeurs de $\tau_A \in \{0.2, 0.4, 0.6\}$ (cf Figures 4.3, 6.9 et 6.10 pour une visualisation de l'évaluation post-prédiction quand $\tau_A = 0.2$, $\tau_A = 0.4$ et $\tau_A = 0.6$ respectivement). Nous indiquons aussi les estimations concernant le paramètre $R_{0,u}$, s'obtenant en fonction de λ_u, d_I, α et τ_A en utilisant l'expression (4.2).

TABLE 4.4 – Estimations de la moyenne, des 5ème et 95ème percentiles et du coefficient de variation des paramètres du modèle $(\lambda_u, R_{0,u}, \tau_{H,u}, \tau_{D,u}, i_{0,u}, \rho_{I,u})^t$ avec $d_H = 20$ jours et en testant trois valeurs de τ_A : 0.2, 0.4 et 0.6. Pour $i_{0,u}$, les valeurs présentées sont calculées sur $1,000,000$ d'habitants.

		λ_u	$R_{0,u}$	$\tau_{H,u}.10^2$	$\tau_{D,u}$	$i_{0,u}.10^6$	$\rho_{I,u}$
Moyenne estimée	$\tau_A = 0.2$	0.71	3.90	0.18	0.12	1551	0.83 %
	$\tau_A = 0.4$	0.78	3.86	0.24	0.13	1174	1.10 %
	$\tau_A = 0.6$	0.87	3.81	0.36	0.13	797	1.69 %
[5, 95]ème percentiles estimés	$\tau_A = 0.2$	[0.59,0.85]	[3.23,4.66]	[0.04,0.44]	[0.08,0.18]	[665,2958]	[0.17,2.20] %
	$\tau_A = 0.4$	[0.64,0.95]	[3.12,4.69]	[0.06,0.58]	[0.08,0.18]	[491,2265]	[0.25,2.84] %
	$\tau_A = 0.6$	[0.73,1.03]	[3.18,4.51]	[0.09,0.89]	[0.08,0.18]	[361,1470]	[0.36,4.40] %
CV estimé	$\tau_A = 0.2$	11 %	11 %	80 %	25 %	48 %	90 %
	$\tau_A = 0.4$	12 %	12 %	79 %	25 %	49 %	84 %
	$\tau_A = 0.6$	11 %	11 %	79 %	25 %	45 %	85 %

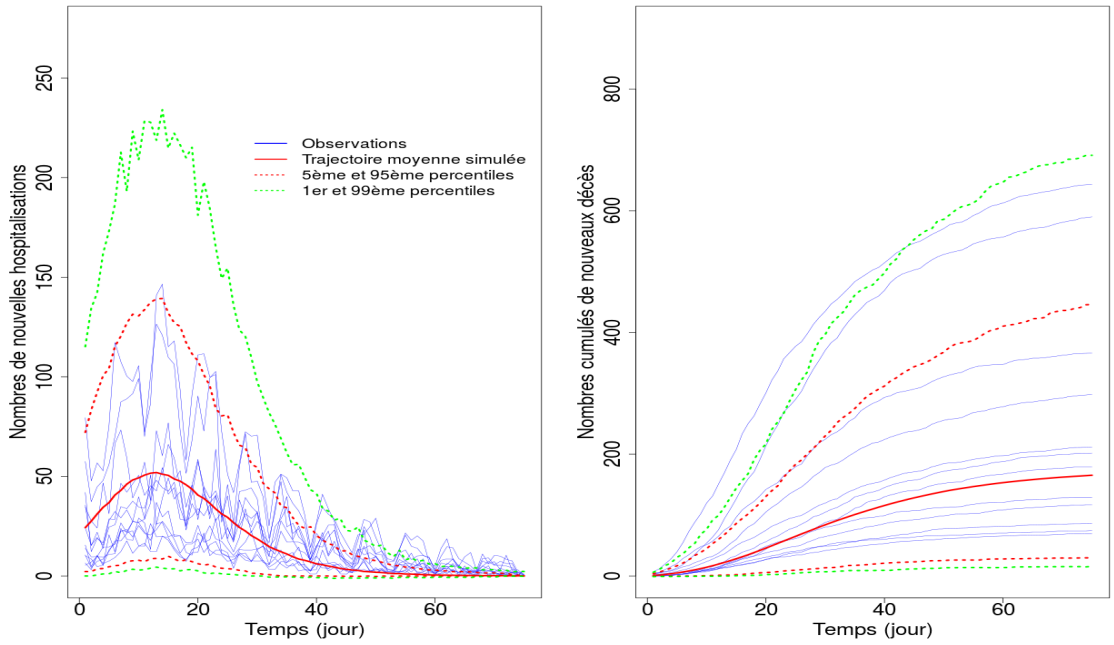

FIGURE 4.3 – Visualisation de l'évaluation post-prédiction avec $\tau_A = 0.2$ et $d_H = 20$ jours. Observations : nombres quotidiens de nouvelles hospitalisations (gauche) et nombres cumulés de nouveaux décès (droite) calculés sur $1,000,000$ d'habitants pour chacune des U épidémies (bleu). Trajectoires simulées obtenues en trois étapes : (i) génération de 1000 valeurs de $\hat{\phi}_u$ en fonction des valeurs estimées des paramètres du modèle ; (ii) sachant $\hat{\phi}_u$, simulation de 1000 épidémies (processus Markoviens de sauts) selon le modèle épidémique ; (iii) calcul de la trajectoire moyenne (ligne pleine rouge), des 5ème et 95ème percentiles (pointillé rouge) et des 1er et 99ème percentiles (pointillé vert) sur les 1000 épidémies simulées.

Pour des raisons d'identifiabilité, nous avons testé plusieurs valeurs de paramètres pour la fraction d'asymptomatiques $\tau_A \in \{0.2, 0.4, 0.6\}$ et de la durée d'hospitalisation $d_H \in \{10, 15, 20\}$ jours. En calculant la log-vraisemblance estimée de chaque modèle (cf Tableau 4.3), nous avons choisi de fixer la durée d'hospitalisation à $d_H = 20$ jours, ce qui est en accord avec celle fixée par Collin et al. (2021) (18.3 jours, cf Tableau 6.8). Concernant les trois valeurs testées du paramètre τ_A, la différence entre les log-vraisemblances estimées est très mineure, avec un maximum atteint en $\tau_A = 0.2$ quelle que soit la valeur de d_H (en cohérence avec la valeur 0.156 fixée par Collin et al. (2021) et la valeur estimée par région, obtenue par Prague et al. (2020), se situant entre 0.17 et 0.19 ; cf Tableau 6.8). Dans le cas $d_H = 20$ jours et $\tau_A = 0.2$, la valeur moyenne estimée du taux de transmission λ (cf Tableau 4.4) est égale à 0.71 (fixée à 0.76 par Cazelles et al. (2021) et estimée à 0.78 par Collin et al. (2021) ; cf Tableau 6.8). La valeur moyenne du ratio de reproduction de base R_0 est quant à elle approximativement estimée à 3.9, entre 3.23 et 4.66 (5ème et 95ème percentiles) selon la région considérée. Ces deux paramètres sont assez variables à l'échelle régionale (CV estimé à 11 %). A titre de comparaison, concernant le paramètre λ, Cazelles et al. (2021) ont fixé différentes valeurs pour les 5 régions considérées (IDF, PACA, OC, NA, ARA ; cf Tableau 6.8), cela impliquant un coefficient de variation égal à 11 %, et Collin et al. (2021) ont estimé un coefficient de variation égal à seulement 1.5 % en considérant 12 régions. La valeur moyenne estimée de la fraction de décès τ_D, proche de 0.12, est légèrement supérieure à celle estimée par Cazelles et al. (2021) (0.042 ; cf Tableau 6.8), tandis que celle de la fraction d'hospitalisés τ_H est estimée à une valeur plus faible, soit moins de 0.01 (Cazelles et al. (2021) : valeur estimée à 0.025 ; Collin et al. (2021) : valeur fixée à 0.034 ; cf Tableau 6.8). Ces paramètres sont assez variables (CV à 80 % pour τ_H vs. 12 % chez Cazelles et al. (2021) ; CV à 25 % pour τ_D vs. 12 % chez

Cazelles et al. (2021)). Dans notre cas, le taux de report moyen du nombre de nouvelles infections ρ_I est estimé à environ 1 %, avec une variabilité entre régions notable (CV estimé supérieur à 90 %). Enfin, la visualisation post-prédiction (cf Figure 4.3) montre que la majorité des observations se situent dans l'enveloppe prédictive (5ème et 95ème percentiles pour les nombres de nouvelles hospitalisations, 1er et 99ème percentiles pour les nombres cumulés de nouveaux décès).

4.6 Discussion

Dans ce chapitre, nous avons analysé des données d'incidence quotidiennes d'infections, d'hospitalisations et de mortalité de la Covid-19 sur 12 régions de la France métropolitaine entre le 28 février et le 1er juin 2020 en utilisant le modèle mécaniste à compartiments SEIRAHD. Pour chaque région $u = 1, \ldots, 12$, les dynamiques épidémiques sont modélisées par un modèle à espace d'état Gaussien et linéaire, celui-ci décrivant une variabilité intra-épidémie. De plus, les paramètres du modèle sont considérés aléatoires à l'échelle de la région afin de prendre en compte explicitement une variabilité inter-épidémies. L'approche d'inférence proposée, couplant l'algorithme SAEM avec des techniques de filtrage de Kalman, est évaluée dans un premier temps sur des données simulées. Les résultats obtenus sont tout à fait satisfaisants, avec des biais d'estimation et écarts-type faibles. L'approche est alors appliquée aux données régionales de la Covid-19.

Ici, l'approche d'inférence a été appliquée sur des données d'incidence. Cependant, sous l'hypothèse assez plausible que le recueil des données de nouvelles hospitalisations et de nouveaux décès est peu sujet à des erreurs de report et connaissant la proportion initiale d'individus hospitalisés et décédés ($h_0 = d_0 = 0$), une alternative serait de s'intéresser à des données cumulées, ce qui revient à appliquer notre méthode d'inférence sur des données de prévalence plutôt que d'incidence. En particulier, Collin et al. (2021) ont considéré à la fois les nombres de nouvelles hospitalisations mais aussi les nombres totaux d'hospitalisations.

Enfin, du point de vue de l'approche d'inférence par des modèles à effets mixtes, le nombre d'individus statistiques (*i.e.* régions ici) considérés est relativement faible ($U = 12$). Pour améliorer la précision des estimations, une perspective serait d'augmenter le nombre d'épidémies à considérer dans la procédure d'inférence. On pourrait par exemple considérer des données à l'échelle départementale ($U \approx 90$) plutôt qu'à l'échelle régionale. Une difficulté est que ces données comportent des valeurs logiquement beaucoup plus faibles, avec une quantité importante de reports de zéro cas par intervalle d'observations dans plusieurs départements. Cela est encore plus marqué concernant les reports de nouveaux décès. Par conséquent, une piste intéressante serait d'adapter la modélisation des observations en s'intéressant à des distributions *zéro-inflatées* (voir e.g. Khedhiri (2021), qui ont montré que prendre en compte une inflation de zéro dans le modèle permet un meilleur ajustement des courbes de mortalité, avec application sur des données quotidiennes de mortalité due à la Covid-19 en Tunisie).

Chapitre 5

Conclusion et perspectives

Le cadre de cette thèse est l'estimation des paramètres de dynamiques épidémiques à partir des données disponibles en cas d'épidémies majeures. Une des retombées est de pouvoir fournir des prédictions les plus fiables possibles de ces dynamiques. Le développement d'approches d'inférence à partir des observations de suivi épidémique est un premier enjeu important, qui est cependant difficile du point de vue statistique. En effet, les données sont généralement recueillies à des temps discrets et sujettes à des erreurs de report et/ou de mesure. A cela s'ajoute la présence de composantes non-observées dans les modèles mécanistes servant à décrire les épidémies, rendant nécessaire l'utilisation d'algorithmes pour l'inférence. De plus, les dynamiques d'une même épidémie peuvent être observées dans des zones géographiques distinctes ou à des périodes différentes. La prise en compte explicite de la variabilité inter-épidémies dans la modélisation et son estimation rigoureuse constituent un deuxième enjeu majeur.

Dans cette thèse, nous avons décrit les dynamiques épidémiques à travers une approche basée sur l'approximation de processus Markoviens de sauts densité-dépendants par des processus Gaussiens à petite variance. En considérant également une approximation Gaussienne du modèle des observations, nous nous sommes plaçés dans le cadre des modèles à espace d'état Gaussiens et linéaires en les états du système, bien que non-linéaires en les paramètres. Via des techniques de filtrage de Kalman, nous avons proposé une méthode d'estimation paramétrique basée sur le calcul d'une vraisemblance approchée des observations. Puis, dans le cas d'épidémies se propageant sur plusieurs sites géographiques simultanément et/ou de façon récurrente sur plusieurs périodes de temps, nous avons utilisé le cadre des modèles à effets mixtes pour décrire de façon plus fine et originale les différentes sources de variabilité des données épidémiques associées. Dans ces modèles, des distributions sont spécifiées pour les paramètres afin de prendre en compte la variabilité inter-épidémies. Pour estimer les paramètres impliqués dans ces distributions, nous avons développé une approche d'inférence combinant des techniques de filtrage de Kalman pour calculer la vraisemblance des observations du modèle et l'algorithme SAEM pour en estimer les paramètres. Dans un premier temps, nous avons évalué les différentes méthodes développées dans cette thèse sur des données simulées d'épidémies décrites par des processus Markoviens de sauts, puis nous les avons appliquées à des données réelles d'incidence quotidienne de syndromes grippaux entre 1990 et 2017 en France et de la Covid-19 dans 12 régions de la France métropolitaine pendant le printemps 2020. En particulier, nous avons contribué à proposer une estimation plus précise des paramètres du modèle et de la variabilité inter-épidémies à l'échelle de l'année pour la grippe humaine et à l'échelle de la région pour la Covid-19.

Des extensions directes du cadre de modélisation et d'inférence développé dans cette thèse sont possibles. Nous en évoquons certaines ci-dessous :

- Nous avons opté pour une description des dynamiques épidémiques par des processus Markoviens de sauts avant d'en considérer des approximations par des processus Gaussiens. Or, l'hypothèse sous-jacente à ce cadre Markovien est que les durées de séjour dans chaque compartiment suivent une distribution exponentielle. Il serait tout à fait possible de considérer des temps de séjour non exponentiels dans certains compartiments via l'inclusion de sous-compartiments avec des temps de séjour distribués exponentiellement. Dans ce cas, la durée totale au sein d'un compartiment correspond à la somme des durées de séjour dans chaque sous-compartiment et suit donc une distribution d'Erlang. Dans notre cadre, la modification s'opère facilement en définissant un modèle avec plus d'états. Cependant, le gain obtenu vis-à-vis de la modélisation est à confronter aux difficultés à pouvoir bien estimer les paramètres du modèle.
- Des modèles plus complexes peuvent être envisagés, comme l'utilisation d'une structure en âge (e.g. Covid-19 : Roux et al. (2021), Di Domenico et al. (2020)) ou la prise en compte de mesures de contrôle telles que la vaccination (e.g. Covid-19 : Collin et al. (2021). Dans ce cas, la dimension du système d'état augmente.
- On pourrait introduire une dépendance temporelle dans les paramètres de la dynamique épidémique afin de prendre en compte une stochasticité environnementale. A titre d'exemple, dans le cas du taux de transmission, une stratégie possible est de le modéliser de façon déterministe par l'introduction d'un forçage saisonnier de la forme

$$\lambda(t) = \lambda_0(1 + \lambda_1 \cos(\omega t)), \tag{5.1}$$

  où λ_0 correspond à un taux de transmission moyen, ω est la période du forçage sinusoïdal et λ_1 décrit l'amplitude de la saisonnalité dans la transmission. Dans ce cas, la dimension de l'espace paramétrique est plus grande.
- L'approche à effets mixtes développée dans le **Chapitre 3** peut inclure des covariables exogènes dans l'objectif de décrire certaines caractéristiques spécifiques à une région et/ou période considérée. Par exemple, cela permettrait de prendre en compte des interventions de santé publique ou des recueils de données différents d'une région à une autre.
- Concernant la modélisation des observations, nous avons proposé une façon d'exprimer les données d'incidence (Y_k) comme une fonction des accroissements $(\Delta_k X)$ des états du système entre deux temps consécutifs via un opérateur de projection $B(\cdot)$:

$$\mathbb{E}(Y_k|\Delta_k X = \Delta_k x) = B\Delta_k x,\ k = 0, \dots, n. \tag{5.2}$$

  Par exemple, dans le cadre du modèle SEIRAHD du **Chapitre 4**, nous avons considéré que le nombre de nouvelles hospitalisations au temps t_k correspond à

$$\int_{t_{k-1}}^{t_k} \tau_H \gamma I(t)\, dt = \Delta_k H + \frac{1}{\tau_D}\Delta_k D,$$

  et conditionnellement à $\Delta_k X = \Delta_k x$, il s'écrit sous la forme (5.2) avec $B = (0\ \ 0\ \ 0\ \ 0\ \ 1\ \ \frac{1}{\tau_D})$ et $\Delta_k X = (\Delta_k S, \Delta_k E, \Delta_k I, \Delta_k A, \Delta_k H, \Delta_k D)^t$. D'une manière plus générale, on pourrait définir un changement de variable directement dans le système d'état en introduisant une variable

$$Z(t) = H(t) + \frac{1}{\tau_D} D(t) \implies H(t) = Z(t) - \frac{1}{\tau_D} D(t),$$

  et en considérant un nouveau système d'état $\Delta_k \tilde{X} = (\Delta_k S, \Delta_k E, \Delta_k I, \Delta_k A, \Delta_k Z, \Delta_k D)^t$. Afin de pouvoir appliquer notre approche d'inférence développée dans cette thèse, il suffirait de

ré-écrire certaines quantités telles que la fonction de dérive et la matrice de diffusion. Par exemple, la fonction de dérive vérifierait dans ce cas :

$$b : (s, e, i, a, z, d)^t \to \begin{pmatrix} -\lambda s(i + \alpha a) \\ \lambda s(i + \alpha a) - \epsilon e \\ (1 - \tau_A)\epsilon e - \gamma i \\ \tau_A \epsilon e - \gamma a \\ \tau_H \gamma i \\ \tau_D \kappa z - \kappa d \end{pmatrix}.$$

- Dans le cas où plusieurs populations sont reliées par des flux migratoires d'individus, on pourrait introduire une approche par métapopulation (e.g. dynamiques régionales connectées ; cf **Introduction**, Section 1.1.1.6). La dimension du système compartimental reste inchangée tandis que la dimension paramétrique augmente (bien que les paramètres associés aux déplacements entre populations peuvent être connus). En notant ϕ_u les paramètres du modèle à compartiments, ce cadre est compatible avec celui proposé dans le **Chapitre 3** :

$$\begin{cases} \phi_u & = h(\beta, \xi_u), \\ \xi_u & \sim \mathcal{N}(0, \Gamma), \end{cases}$$

  avec $h(\cdot)$ une fonction de lien connue et β le vecteur d'effets fixes. La différence vient dans la modélisation des effets aléatoires (ξ_u). En effet, comme les épidémies locales ne sont plus supposées indépendantes, cela induit des corrélations entre les effets aléatoires et donc la matrice de variance Γ n'est plus diagonale. Une conséquence directe est l'augmentation de l'espace paramétrique et peut à la fois occasionner des coûts de calcul plus importants et des difficultés à bien estimer tous les paramètres à partir des données disponibles.

Une autre catégorie de perspectives pourrait se traiter dans le même cadre méthodologique que celui de cette thèse, mais les développements théoriques ne sont pas immédiats :

- Nous avons évoqué précédemment la possibilité de considérer des paramètres non homogènes dans le temps (cf (5.1)). Une première alternative serait d'utiliser une modélisation non-paramétrique de ces paramètres. Par exemple, on pourrait supposer que le taux de transmission $\lambda(t)$ est une fonction déterministe inconnue que l'on chercherait ensuite à estimer à partir des données disponibles.
- Une deuxième alternative serait de modéliser certains paramètres par des processus stochastiques puis estimer les paramètres régissant ces processus à partir des données disponibles. Par exemple, Cazelles et al. (2021) ont supposé un taux de transmission $\lambda(t)$ suivant un processus de diffusion :

$$d \log(\lambda(t)) = \nu . dB(t),$$

  où ν est la volatilité du processus Brownien (dB). La valeur de ν est inconnue et serait donc à estimer.

Enfin, nous pouvons évoquer des questions ouvertes qui nécessiteraient le développement de méthodes spécifiques :

- La cadre développé dans cette thèse repose sur la description des dynamiques d'un modèle à compartiments générique par des processus Markoviens de sauts, ce qui implique que les durées de séjour dans chaque compartiment suivent des distributions exponentielles. Pour relâcher cette hypothèse forte, on pourrait envisager d'utiliser des processus non-Markoviens

pour modéliser les dynamiques épidémiques. Par exemple, cela permettrait de prendre en compte une durée de guérison dépendant du temps écoulé depuis l'infection. Ainsi, une perspective serait de développer un cadre théorique pour l'estimation des paramètres épidémiques.

- Enfin, dans cette thèse, nous avons toujours considéré des approximations Gaussiennes des modèles des observations afin de poser un cadre statistique fondé sur les modèles à espace d'état. Par exemple, la distribution Gaussienne peut être choisie pour approximer les moyenne et variance d'une distribution binomiale :

$$Y_k|(X_k;\mu) \sim \mathcal{N}\left(pX_k, p(1-p)X_k\right),\ \mu = p,$$

avec (Y_k) les observations, (X_k) les états cachés et p le taux de report. En particulier, cette approximation peut être particulièrement mauvaise lorsque X_k est petit. Une alternative serait donc de considérer le modèle d'observation original, soit :

$$Y_k|(X_k;\mu) \sim \text{Binomiale}\left(X_k, p\right),\ \mu = p.$$

Dans cette situation, il n'est plus possible d'écrire une équation d'état de la forme

$$Y_k|(X_k;\mu) = B(\mu)X_k + V_k,$$

avec (V_k) un bruit Gaussien quelconque. La conséquence est qu'on ne peut plus écrire le modèle statistique sous le formalisme des modèles à espace d'état linéaires et Gaussiens, et *a fortiori* utiliser des techniques de filtrage de Kalman linéaire. Cela nécessiterait donc le développement d'algorithmes récursifs pour calculer la vraisemblance des observations du modèle.

Chapitre 6

Annexes

Table des matières

A Annexes : Chapitre 2 108
A.1 Remarks on the sampling interval 108
A.2 Proof of Proposition 2 108
A.3 Proof of Lemma 2 109
A.4 Proof of Proposition 3 109
A.5 Additional simulation study 109
A.6 User-friendly code 116
B Annexes : Chapitre 3 117
B.1 Key quantities involved in the SEIR epidemic model 117
B.2 Details on the Kalman filter equations for incidence data of epidemic dynamics 117
B.3 Practical considerations on implementation setting 118
B.4 Estimation results for a second set of parameter values 119
C Annexes : Chapitre 4 123
C.1 Modèles mécanistes utilisés dans la littérature pour décrire la propagation du SARS-CoV-2 dans une population en France 123
C.2 Tableau résumé de plusieurs études d'inférence 126
C.3 Visualisation de l'évaluation post-prédictive pour $\tau_A = 0.4$ et $\tau_A = 0.6$. 128

A Annexes : Chapitre 2

A.1 Remarks on the sampling interval

The sampling interval Δ is important in our method and we distinguish between two cases : "Small Δ" and "Moderate Δ". We give below the dependencies on quantities of interest with respect to Δ.

(1) Small sampling interval Δ
Taylor expansions with respect to t at point t_{k-1} yield

$$\begin{aligned} F_k(\eta) &= F_k(\eta, \Delta) = \Delta\left(b(\eta, x(\eta, t_{k-1})) - \nabla_x b(\eta, x(\eta, t_{k-1}))x(\eta, t_{k-1})\right) + \Delta\, o(1), \\ A_k(\eta) &= A_k(\eta, \Delta) = I_d + \Delta \nabla_x b(\eta, x(\eta, t_{k-1})) + \Delta\, o(1), \\ T_k(\eta) &= T_k(\eta, \Delta) = \frac{1}{N}\left(\Delta\Sigma(\eta, x(\eta, t_{k-1})) + \Delta\, o(1)\right). \end{aligned}$$

The following additional approximations, which simplify the analytic expressions, can be used in the state space equation :

$$\begin{aligned} &X_k = \Delta\left(b(\eta, x(\eta, t_{k-1})) - \nabla_x b(\eta, x(\eta, t_{k-1}))x(\eta, t_{k-1})\right) + \left(I_d + \Delta \nabla_x b(\eta, x(\eta, t_{k-1}))\right) X_{k-1} + U_k, \\ &U_k \sim \mathcal{N}_d\left(0, \frac{\Delta}{N}\Sigma(\eta, x(\eta, t_{k-1}))\right). \end{aligned}$$

(2) Moderate Δ
Computing the approximate log-likelihood (2.20) with Kalman filtering techniques requires computing the resolvent matrix Φ of the ODE system (2.8). When the time intervals between observations are too large (i.e., Δ is too large), we use the following approximation for matrix exponentials :

$$\Phi\left(\theta_x, t_{k+1}, t_k\right) \approx \prod_{j=1,\ldots,J-1} \left(I_d + (a_{j+1} - a_j)\nabla_x b(\theta_x, x(\theta_x, a_j))\right), \tag{6.1}$$

where $t_k = a_1 < a_2 < \ldots < a_J = t_{k+1}$. This can however significantly increase computation times.

A.2 Proof of Proposition 2

By the semigroup property of Φ, we have that g, defined in (2.7), satisfies for $s \leq t$,

$$\begin{aligned} g(t) &= \Phi(t, s)\int_0^s \Phi(s, u)\sigma(x(u))dB(u) + \int_s^t \Phi(t, u)\sigma(x(u))dB(u), \\ &= \Phi(t, s)g(s) + \int_s^t \Phi(t, u)\sigma(x(u))dB(u). \end{aligned}$$

Substituting $g(s)$ with $\sqrt{N}(G_N(s) - x(s))$ using (2.9) yields :

$$G_N(t) = x(t) + \Phi(t, s)(G_N(s) - x(s)) + \frac{1}{\sqrt{N}}\int_s^t \Phi(t, u)\sigma(x(u))dB(u).$$

Setting $F(t_k) = x(t_k) - \Phi(t_k, t_{k-1})x(t_{k-1})$ and $U_k = \int_{t_{k-1}}^{t_k} \Phi(t_k, u)\sigma(x(u))dB(u)$ yields (ii). Clearly, U_k is $\mathcal{F}_{t_k}$-measurable. By the independent increments property of Brownian motion, we get moreover that U_k is independent of $\mathcal{F}_{t_{k-1}}$. This achieves the proof of Proposition 2.

A.3 Proof of Lemma 2

Assume first that Q and T are non-singular. The joint distribution of (Y, X) is Gaussian :

$$\mathcal{L}(Y,X) \simeq \exp\{-\frac{1}{2}\Big((y-Bx)^t Q^{-1}(y-Bx) + (x-\xi)^t T^{-1}(x-\xi)\Big)\}.$$

Hence,

$$\mathcal{L}(X|Y) \simeq \exp\{-\frac{1}{2}\Big(x^t(B^tQ^{-1}B+T^{-1})x - 2x^t(B^tQ^{-1}y+T^{-1}\xi)\Big)\}.$$

Setting

$$\bar{T} = (B^tQ^{-1}B+T^{-1})^{-1} = (I_d + TB^tQ^{-1}B)^{-1}T,$$

we get :

$$\mathcal{L}(X|Y) \simeq \exp\{-\frac{1}{2}\Big((x-\bar{T}(B^tQ^{-1}y+T^{-1}\xi))^t\bar{T}^{-1}(x-\bar{T}(B^tQ^{-1}y+T^{-1}\xi))\Big)\},$$

and

$$\bar{\xi}(y) = (I_d + TB^tQ^{-1}B)^{-1}T(T^{-1}\xi + B^tQ^{-1}y) = (I_d + TB^tQ^{-1}B)^{-1}(\xi + TB^tQ^{-1}y).$$

We then obtain, using the matrix relation :

$$(I_d + TB^tQ^{-1}B)^{-1} = I_d - TB^t(BTB^t+Q)^{-1}B,$$

the following results :

$$\begin{aligned}
\bar{\xi}(y) &= \xi - TB^t(BTB^t+Q)^{-1}B\xi + TB^t(Q^{-1} - (BTB^t+Q)^{-1}TB^tQ^{-1})y, \\
&= \xi + TB^t(BTB^t+Q)^{-1}(y-B\xi), \\
\bar{T} &= (I_d + TB^tQ^{-1}B)^{-1}T = T - TB^t(BTB^t+Q)^{-1}BT.
\end{aligned}$$

A.4 Proof of Proposition 3

For $k = 0$, we have that $X_0 \sim \mathcal{N}(\xi_0, \hat{\Xi}_0)$. The induction assumption is : $\mathcal{L}(X_k|Y_{k-1,0}) = \mathcal{N}_d(\hat{X}_k, \hat{\Xi}_k)$, with $k \geq 1$.

To get (i), we apply Lemma 2, noting that the distribution $\mathcal{L}(X_k|Y_{k-1,0}) = \mathcal{N}_d(\hat{X}_k, \hat{\Xi}_k)$ and that the distribution Y_k conditional on X_k is $\mathcal{N}(BX_k, Q_k)$. Therefore, setting $\xi = \hat{X}_k$, $T = \hat{\Xi}_k$ and $Q = Q_k$, we get that the distribution of $(X_k|Y_{k:0})$ is $\mathcal{N}_d(\bar{X}_k, \bar{T}_k)$, with $\bar{X}_k = \bar{\xi}(Y_k)$, where $\bar{\xi}(Y_k)$ and $\bar{T}_k$ are given by (2.21). These are precisely the expressions for $\bar{T}_k$ and $\bar{\Xi}_k$ given in (i).

For (ii), we use that $X_{k+1} = F_{k+1} + A_kX_k + U_{k+1}$ and $\mathcal{L}(X_k|Y_{k:0}) \sim \mathcal{N}_d(\bar{X}_k, \bar{T}_k)$. Therefore, $\mathcal{L}(X_{k+1}|Y_{k:0}) = \mathcal{N}_d(F_{k+1} + A_k\bar{X}_k, A_k\bar{T}_kA_k^t + T_{k+1})$. Setting $\hat{X}_{k+1} = F_{k+1} + A_k\bar{X}_k$ and $\hat{\Xi}_{k+1} = A_k\bar{T}_kA_k^t + T_{k+1}$ yields (ii).

For (iii), we use that $Y_{k+1} = BX_{k+1} + V_{k+1}$ and that $\mathcal{L}(X_{k+1}|Y_{k:0}) \sim \mathcal{N}(\hat{X}_{k+1}, \hat{\Xi}_{k+1})$. This gives that $\mathcal{L}(Y_{k+1}|Y_{k:0})$ is equal to $\mathcal{N}_q(B\hat{X}_{k+1}, B\hat{\Xi}_{k+1}B^t + Q_{k+1})$. Setting $\hat{M}_{k+1} = B\hat{X}_{k+1}, \hat{\Omega}_{k+1} = B\hat{\Xi}_{k+1}B^t + Q_{k+1}$ yields (iii). The induction assumption is fulfilled and therefore this achieves the proof of Proposition 3.

A.5 Additional simulation study

A.5.1 Description

We reproduced the simulation study described in Section 2.4 with other parameter values : $\lambda = 0.6$, $\gamma = 0.4$, $s_0 = 0.99$, $i_0 = 0.01$. An extract of the simulated data is shown in Figure 6.1.

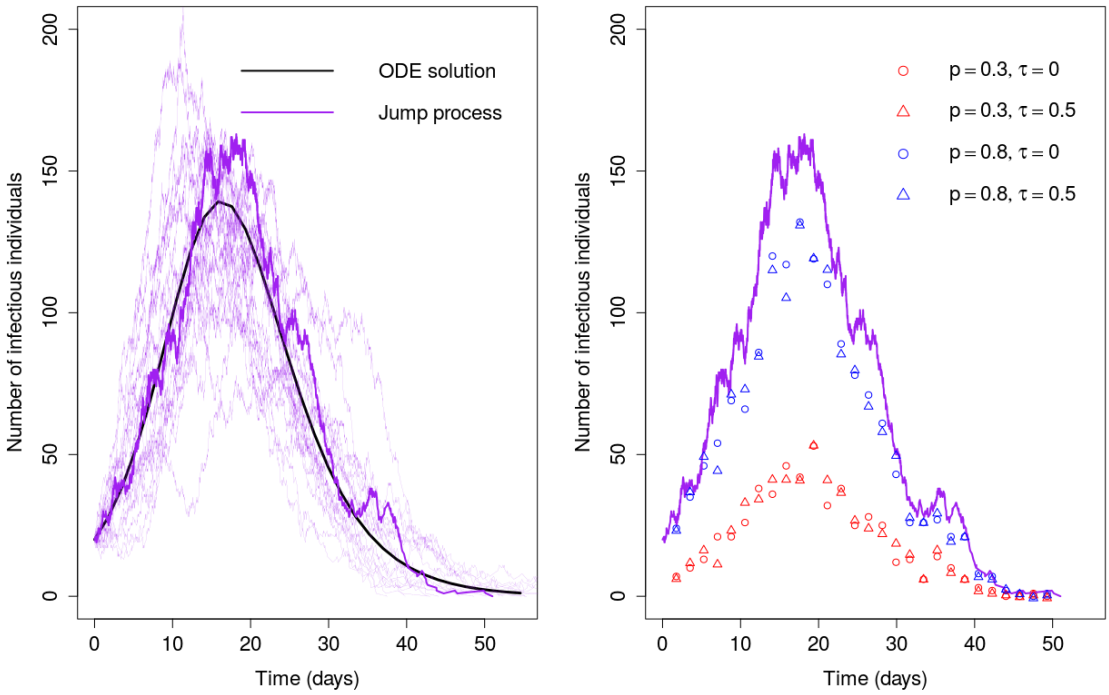

FIGURE 6.1 – Left panel : ODE solution for the number of infected individuals I (plain black line) and 20 trajectories of the Markov jump process for I (purple lines) when $N = 2000$. Right panel : $n = 30$ observations obtained from a particular trajectory of the jump process (in bold purple in the left panel) as a function of time. The points and triangles stand for observations generated with measurement error terms $\tau = 0$ and $\tau = 0.5$ respectively, and the blue and red symbols represent observations generated with $p = 0.8$ and $p = 0.3$ respectively.

A.5.2 Point estimates and standard deviations for key model parameters θ

Numerical results for the first experiment ($\tau = 0$) Tables 6.1 and 6.2 respectively display the results for the high-reporting scenario ($p = 0.8$) and low reporting scenario ($p = 0.3$) when $\tau = 0$ and is not estimated. Each table compares the Kalman-based method (KM) to the maximum iterated filtering algorithm (MIF). The first column display the true parameter values. Columns 2 to 10 display the results for different combinations of (N, n). For each parameter and each estimation method, the reported values are the mean of the 500 parameter estimates and their standard deviations (in brackets).

TABLE 6.1 – First experiment ($\tau = 0$). Estimation of $\theta = (\lambda, \gamma, p, i_0)$ under the constraint $s_0 + i_0 = 1$ in Setting 1 with true parameter values $(\lambda^*, \gamma^*, p^*, i_0^*)$=(0.6, 0.4, 0.8, 0.01). For each combination of (N, n) and for each model parameter, point estimates and standard deviations are calculated as the mean of the 500 individual estimates and their standard deviations (in brackets) obtained by our Kalman-based method (KM) and maximum iterated filtering (MIF). The reported values for the number of observations n correspond to the average over the 500 trajectories, with the min and max in brackets.

	$N = 1000$			$N = 2000$			$N = 10000$		
	$n = 10$	$n = 30$	$n = 99$	$n = 11$	$n = 31$	$n = 102$	$n = 11$	$n = 31$	$n = 101$
	(3, 19)	(9, 56)	(30, 182)	(7, 19)	(20, 56)	(66, 182)	(8, 17)	(24, 49)	(78, 160)
$\lambda^* = 0.6$									
KM	0.47	0.50	0.59	0.45	0.50	0.59	0.48	0.51	0.60
	(0.16)	(0.15)	(0.16)	(0.08)	(0.10)	(0.08)	(0.04)	(0.06)	(0.05)
MIF	0.50	0.53	0.58	0.49	0.52	0.59	0.51	0.52	0.60
	(0.14)	(0.17)	(0.11)	(0.09)	(0.09)	(0.07)	(0.06)	(0.06)	(0.04)
$\gamma^* = 0.4$									
KM	0.19	0.27	0.39	0.21	0.28	0.39	0.25	0.29	0.40
	(0.08)	(0.11)	(0.09)	(0.07)	(0.09)	(0.04)	(0.05)	(0.07)	(0.03)
MIF	0.22	0.30	0.39	0.25	0.31	0.40	0.29	0.32	0.41
	(0.11)	(0.10)	(0.06)	(0.10)	(0.08)	(0.04)	(0.07)	(0.07)	(0.03)
$p^* = 0.8$									
KM	0.28	0.49	0.75	0.32	0.50	0.77	0.41	0.50	0.82
	(0.20)	(0.27)	(0.11)	(0.18)	(0.22)	(0.08)	(0.13)	(0.18)	(0.09)
MIF	0.37	0.53	0.78	0.40	0.55	0.78	0.49	0.55	0.83
	(0.24)	(0.24)	(0.07)	(0.22)	(0.21)	(0.07)	(0.17)	(0.18)	(0.07)
$i_0^* = 0.01$									
KM	0.029	0.032	0.013	0.025	0.022	0.012	0.018	0.019	0.011
	(0.034)	(0.081)	(0.019)	(0.021)	(0.013)	(0.005)	(0.006)	(0.006)	(0.003)
MIF	0.027	0.025	0.012	0.023	0.019	0.011	0.016	0.016	0.010
	(0.029)	(0.048)	(0.003)	(0.019)	(0.011)	(0.002)	(0.006)	(0.006)	(0.001)

TABLE 6.2 – First experiment ($\tau = 0$). Estimation of $\theta = (\lambda, \gamma, p, i_0)$ under the constraint $s_0 + i_0 = 1$ in Setting 2 with true parameter values $(\lambda^*, \gamma^*, p^*, i_0^*)$=(0.6, 0.4, 0.3, 0.01). For each combination of (N, n) and for each model parameter, point estimates and standard deviations are calculated as the mean of the 500 individual estimates and their standard deviations (in brackets) obtained by our Kalman-based method (KM) and maximum iterated filtering (MIF). The reported values for the number of observations n correspond to the average over the 500 trajectories, with the min and max in brackets.

	$N = 1000$			$N = 2000$			$N = 10000$		
	$n = 10$	$n = 30$	$n = 99$	$n = 11$	$n = 31$	$n = 102$	$n = 11$	$n = 31$	$n = 101$
	(3, 19)	(9, 56)	(30, 182)	(7, 19)	(20, 56)	(66, 182)	(8, 17)	(24, 49)	(78, 160)
$\lambda^* = 0.6$									
KM	0.44	0.50	0.53	0.43	0.47	0.54	0.48	0.50	0.55
	(0.18)	(0.12)	(0.15)	(0.08)	(0.09)	(0.09)	(0.06)	(0.06)	(0.07)
MIF	0.47	0.51	0.55	0.47	0.49	0.53	0.52	0.53	0.55
	(0.12)	(0.11)	(0.15)	(0.09)	(0.08)	(0.08)	(0.09)	(0.06)	(0.06)
$\gamma^* = 0.4$									
KM	0.17	0.19	0.29	0.17	0.21	0.31	0.26	0.28	0.34
	(0.19)	(0.09)	(0.09)	(0.06)	(0.08)	(0.08)	(0.07)	(0.06)	(0.08)
MIF	0.20	0.21	0.29	0.22	0.24	0.31	0.31	0.31	0.35
	(0.09)	(0.10)	(0.10)	(0.09)	(0.09)	(0.08)	(0.11)	(0.07)	(0.07)
$p^* = 0.3$									
KM	0.08	0.11	0.19	0.08	0.12	0.21	0.16	0.17	0.24
	(0.08)	(0.08)	(0.09)	(0.04)	(0.07)	(0.09)	(0.07)	(0.07)	(0.09)
MIF	0.11	0.12	0.18	0.12	0.13	0.20	0.21	0.20	0.24
	(0.10)	(0.09)	(0.08)	(0.09)	(0.07)	(0.08)	(0.12)	(0.07)	(0.07)
$i_0^* = 0.01$									
KM	0.028	0.020	0.023	0.023	0.019	0.015	0.020	0.017	0.013
	(0.069)	(0.022)	(0.078)	(0.016)	(0.012)	(0.010)	(0.009)	(0.006)	(0.006)
MIF	0.025	0.022	0.023	0.022	0.020	0.015	0.018	0.015	0.013
	(0.027)	(0.029)	(0.061)	(0.016)	(0.015)	(0.009)	(0.009)	(0.005)	(0.005)

The results on the second set of epidemic parameters displayed in Tables 6.1 and 6.2 are more contrasted, since the parameter values chosen ($\lambda^* = 0.6$ and $\gamma^* = 0.4$) generate more stochasticity (see Figure 6.1), so trajectories are less similar and further from the mean of the jump process; hence estimates are less accurate. Besides, the peak of the number of infectious individuals is clearly lower than in the $\lambda^* = 1$ and $\gamma^* = 1/3$ case. The estimates of p are particularly poor when n is low, which obviously impacts estimation of the other parameters.

Numerical results for the second experiment ($\tau \neq 0$)

Unknown starting point i_0 Table 6.3 displays the results obtained by our Kalman-based method and the MIF algorithm for the high-reporting scenario ($p = 0.8$).

TABLE 6.3 – Second experiment ($\tau \neq 0$). Estimation of $\theta = (\lambda, \gamma, p, i_0, \tau)$ under the constraint $s_0 + i_0 = 1$ in Setting 1 with true parameter values $(\lambda^*, \gamma^*, p^*, i_0^*, \tau^*)$=(0.6, 0.4, 0.8, 0.01, 0.5). For each combination of (N, n) and for each model parameter, point estimates and standard deviations are calculated as the mean of the 500 individual estimates and their standard deviations (in brackets) obtained by our Kalman-based method and the MIF algorithm. The reported values for the number of observations n correspond to the average over the 500 trajectories, with the min and max in brackets.

	$N = 1000$					$N = 10000$				
	$n = 10$	$n = 30$	$n = 99$	$n = 499$	$n = 998$	$n = 11$	$n = 31$	$n = 101$	$n = 500$	$n = 1001$
	(3, 19)	(9, 56)	(30, 182)	(152, 916)	(304, 1831)	(8, 17)	(24, 49)	(78, 160)	(385, 789)	(771, 1577)
$\lambda^* = 0.6$										
KM	0.50	0.49	0.56	0.58	0.57	0.48	0.50	0.55	0.57	0.58
	(0.32)	(0.15)	(0.13)	(0.12)	(0.14)	(0.04)	(0.07)	(0.06)	(0.05)	(0.07)
MIF	0.52	0.51	0.56	0.59	0.58	0.52	0.52	0.56	0.58	0.59
	(0.13)	(0.10)	(0.11)	(0.10)	(0.11)	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)
$\gamma^* = 0.4$										
KM	0.20	0.25	0.35	0.39	0.39	0.25	0.28	0.34	0.38	0.39
	(0.33)	(0.13)	(0.09)	(0.07)	(0.09)	(0.04)	(0.07)	(0.06)	(0.03)	(0.04)
MIF	0.25	0.30	0.36	0.39	0.38	0.30	0.32	0.36	0.39	0.40
	(0.07)	(0.07)	(0.08)	(0.07)	(0.07)	(0.05)	(0.05)	(0.05)	(0.04)	(0.04)
$p^* = 0.8$										
KM	0.24	0.42	0.66	0.77	0.78	0.39	0.49	0.65	0.75	0.77
	(0.16)	(0.25)	(0.23)	(0.16)	(0.14)	(0.11)	(0.20)	(0.19)	(0.11)	(0.11)
MIF	0.39	0.50	0.65	0.72	0.72	0.51	0.55	0.68	0.76	0.79
	(0.16)	(0.17)	(0.17)	(0.13)	(0.15)	(0.13)	(0.15)	(0.15)	(0.13)	(0.13)
$i_0^* = 0.01$										
KM	0.029	0.037	0.022	0.016	0.015	0.019	0.017	0.014	0.012	0.011
	(0.048)	(0.086)	(0.078)	(0.039)	(0.010)	(0.006)	(0.007)	(0.004)	(0.003)	(0.004)
MIF	0.014	0.016	0.014	0.012	0.012	0.015	0.015	0.013	0.011	0.011
	(0.005)	(0.005)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.002)	(0.002)
$\tau^* = 0.5$										
KM	0.34	0.44	0.52	0.54	0.52	0.11	0.21	0.35	0.42	0.47
	(0.69)	(0.44)	(0.20)	(0.16)	(0.17)	(0.22)	(0.21)	(0.20)	(0.13)	(0.13)
MIF	0.49	0.49	0.47	0.48	0.43	0.46	0.39	0.35	0.44	0.49
	(0.25)	(0.19)	(0.17)	(0.15)	(0.20)	(0.24)	(0.18)	(0.17)	(0.15)	(0.14)

Known starting point i_0 Tables 6.4 and 6.5 respectively display the results obtained by our Kalman-based method and the MIF algorithm for the high-reporting scenario ($p = 0.8$) and low-reporting scenario ($p = 0.3$).

TABLE 6.4 – Second experiment ($\tau \neq 0$). Estimation of $\theta = (\lambda, \gamma, p, \tau)$ with $s_0 = 0.99$ and $i_0 = 0.01$ known in Setting 1 with true parameter values $(\lambda^*, \gamma^*, p^*, \tau^*)$=(0.6, 0.4, 0.8, 0.5). For each combination of (N, n) and for each model parameter, point estimates and standard deviations are calculated as the mean of the 500 individual estimates and their standard deviations (in brackets) obtained by KM and MIF. The reported values for the number of observations n correspond to the average over the 500 trajectories, with the min and max in brackets.

	$N = 1000$			$N = 2000$			$N = 10000$		
	$n = 10$	$n = 30$	$n = 99$	$n = 11$	$n = 31$	$n = 102$	$n = 11$	$n = 31$	$n = 101$
	(3, 19)	(9, 56)	(30, 182)	(7, 19)	(20, 56)	(66, 182)	(8, 17)	(24, 49)	(78, 160)
$\lambda^* = 0.6$									
KM	0.54	0.57	0.59	0.55	0.58	0.60	0.56	0.56	0.59
	(0.15)	(0.15)	(0.13)	(0.11)	(0.10)	(0.08)	(0.06)	(0.06)	(0.05)
MIF	0.55	0.55	0.59	0.56	0.57	0.60	0.57	0.58	0.60
	(0.12)	(0.10)	(0.11)	(0.09)	(0.07)	(0.07)	(0.04)	(0.03)	(0.03)
$\gamma^* = 0.4$									
KM	0.27	0.35	0.38	0.30	0.36	0.39	0.35	0.36	0.39
	(0.11)	(0.11)	(0.10)	(0.09)	(0.07)	(0.05)	(0.05)	(0.05)	(0.04)
MIF	0.27	0.34	0.39	0.30	0.37	0.40	0.36	0.39	0.41
	(0.09)	(0.07)	(0.07)	(0.09)	(0.05)	(0.04)	(0.04)	(0.03)	(0.02)
$p^* = 0.8$									
KM	0.42	0.61	0.72	0.49	0.67	0.76	0.65	0.69	0.77
	(0.22)	(0.22)	(0.18)	(0.20)	(0.19)	(0.15)	(0.14)	(0.15)	(0.12)
MIF	0.43	0.62	0.75	0.49	0.67	0.78	0.67	0.75	0.80
	(0.19)	(0.17)	(0.13)	(0.19)	(0.14)	(0.10)	(0.12)	(0.09)	(0.07)
$\tau^* = 0.5$									
KM	0.41	0.63	0.60	0.32	0.65	0.62	0.09	0.30	0.46
	(0.47)	(0.33)	(0.20)	(0.42)	(0.25)	(0.16)	(0.26)	(0.28)	(0.16)
MIF	0.52	0.56	0.55	0.55	0.62	0.56	0.50	0.50	0.47
	(0.27)	(0.20)	(0.13)	(0.29)	(0.18)	(0.12)	(0.27)	(0.19)	(0.10)

TABLE 6.5 – Second experiment ($\tau \neq 0$). Estimation of $\theta = (\lambda, \gamma, p, \tau)$ with $s_0 = 0.99$ and $i_0 = 0.01$ known in Setting 2 with true parameter values $(\lambda^*, \gamma^*, p^*, \tau^*)$=(0.6, 0.4, 0.3, 0.5). For each combination of (N, n) and for each model parameter, point estimates and standard deviations are calculated as the mean of the 500 individual estimates and their standard deviations (in brackets) obtained by KM and MIF. The reported values for the number of observations n correspond to the average over the 500 trajectories, with the min and max in brackets.

	$N = 1000$			$N = 2000$			$N = 10000$		
	$n = 10$	$n = 30$	$n = 99$	$n = 11$	$n = 31$	$n = 102$	$n = 11$	$n = 31$	$n = 101$
	(3, 19)	(9, 56)	(30, 182)	(7, 19)	(20, 56)	(66, 182)	(8, 17)	(24, 49)	(78, 160)
$\lambda^* = 0.6$									
KM	0.47	0.57	0.57	0.47	0.54	0.58	0.56	0.55	0.56
	(0.16)	(0.13)	(0.16)	(0.12)	(0.09)	(0.09)	(0.06)	(0.06)	(0.05)
MIF	0.51	0.55	0.55	0.52	0.55	0.56	0.57	0.57	0.59
	(0.12)	(0.13)	(0.11)	(0.09)	(0.07)	(0.07)	(0.05)	(0.04)	(0.03)
$\gamma^* = 0.4$									
KM	0.20	0.29	0.35	0.26	0.30	0.37	0.37	0.35	0.35
	(0.12)	(0.13)	(0.14)	(0.10)	(0.09)	(0.09)	(0.05)	(0.06)	(0.05)
MIF	0.22	0.26	0.32	0.25	0.30	0.34	0.38	0.37	0.39
	(0.08)	(0.10)	(0.09)	(0.10)	(0.08)	(0.06)	(0.06)	(0.04)	(0.03)
$p^* = 0.3$									
KM	0.10	0.19	0.24	0.15	0.20	0.27	0.28	0.25	0.25
	(0.09)	(0.15)	(0.12)	(0.08)	(0.12)	(0.13)	(0.06)	(0.08)	(0.10)
MIF	0.11	0.14	0.19	0.14	0.18	0.21	0.27	0.27	0.28
	(0.06)	(0.09)	(0.06)	(0.08)	(0.07)	(0.05)	(0.07)	(0.04)	(0.03)
$\tau^* = 0.5$									
KM	0.23	0.42	0.52	0.13	0.24	0.51	0.05	0.19	0.36
	(0.20)	(0.29)	(0.21)	(0.19)	(0.21)	(0.18)	(0.14)	(0.15)	(0.12)
MIF	0.38	0.37	0.44	0.39	0.29	0.43	0.61	0.33	0.43
	(0.14)	(0.15)	(0.09)	(0.15)	(0.15)	(0.07)	(0.17)	(0.12)	(0.05)

A.5.3 Numerical confidence intervals

Figures 6.2 and 6.3 represent the profile likelihoods and the subsequent confidence intervals (CI95%) for the parameters λ and γ obtained for our Kalman filtering-based method in two settings (first case : $N = 2000$, $n = 30$, and $p = 0.3$; second case : $N = 10000$, $n = 100$, and $p = 0.8$).

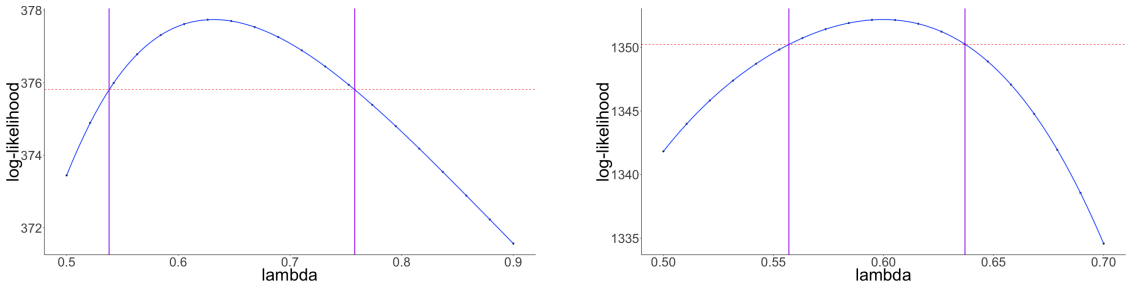

FIGURE 6.2 – Profile likelihood and confidence intervals (CI95%) for λ. Left panel : $N = 2000$, $n = 30$, and $p = 0.3$. The true value $\lambda^* = 0.6$, the point estimate $\hat{\lambda} = 0.47$, and CI95% = [0.54, 0.76]. Right panel : $N = 10000$, $n = 100$, and $p = 0.8$. The true value $\lambda^* = 0.6$, the point estimate $\hat{\lambda} = 0.60$, and CI95% = [0.56, 0.64].

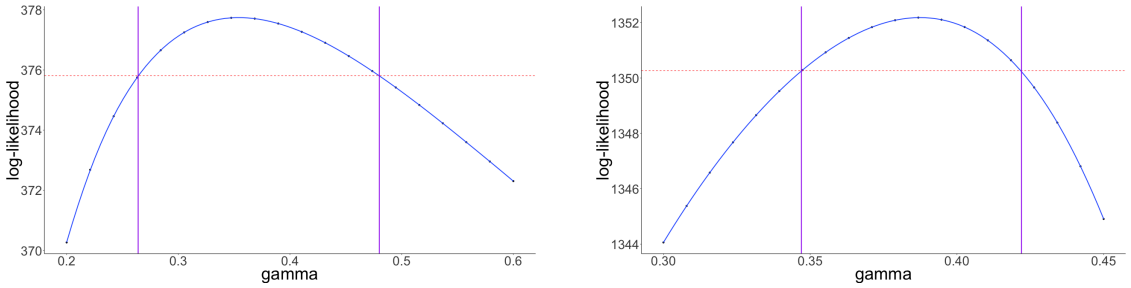

FIGURE 6.3 – Profile likelihood and confidence intervals (CI95%) for γ. Left panel : $N = 2000$, $n = 30$, and $p = 0.3$. The true value $\gamma^* = 0.4$, the point estimate $\hat{\gamma} = 0.21$, and CI95% $= [0.26, 0.48]$. Right panel : $N = 10000$, $n = 100$, and $p = 0.8$. The true value $\gamma^* = 0.4$, the point estimate $\hat{\gamma} = 0.40$, and CI95% $= [0.35, 0.42]$.

A.6 User-friendly code

We propose user-friendly code composed of four distinct programs in the R language, available at the RunMyCode website : `http://www.runmycode.org/companion/view/4074`.

— *KalmanFunctions.R* includes general functions implementing the Kalman filter and computing the likelihood of the observations, given a specified compartmental model, with a fixed sampling interval. These functions are easily generalizable to the case where the sampling interval is variable. Moreover, this script includes a function computing the resolvent matrix defined in (2.8) for large time intervals between observations Δ.

— *ModelFunctions.R* implements the SIR and SEIR models and defines the key quantities (described in the manuscript for the SIR model) necessary to apply the Kalman filter-based method. More precisely, given a compartmental model (SIR or SEIR), the following functions are implemented : the ode system, the drift function, the gradient of the drift function, the diffusion matrix, the projection operator linking the observations to the states of the epidemic model and the variance of the observations.

— *SIRexample.R* and *SEIRexample.R* simulate respectively SIR and SEIR Markovian jump processes for a set of parameters values, using the `GillespieSSA` package. The observations of infectious individuals are obtained by : $O_1(t_k) \sim \text{Binomial}(I(t_k), p)$, $O_2(t_k) \sim \mathcal{N}(0, \tau^2 I(t_k))$, $k = 1, \ldots, n$, at regularly-spaced time points. Finally, an estimation of key parameters λ, γ, p and τ with known starting points and, in the SEIR model, with a known transition rate from E to I, is proposed.

B Annexes : Chapitre 3

B.1 Key quantities involved in the SEIR epidemic model

In the SEIR model, epidemic parameters are the transition rates λ, ϵ and γ and the initial proportions of susceptible, exposed and infectious individuals $s_0 = \frac{S(0)}{N}$, $e_0 = \frac{E(0)}{N}$ and $i_0 = \frac{I(0)}{N}$. When there is no ambiguity, we denote by s, e and i respectively the solutions $s(\eta, t)$, $e(\eta, t)$ and $i(\eta, t)$ of the system of ODEs defined in (3.25). Then, the functions $b(\eta, \cdot)$ and $\Sigma(\eta, \cdot)$ are

$$b(\eta, s, e, i) = \begin{pmatrix} -\lambda si \\ \lambda si - \epsilon e \\ \epsilon e - \gamma i \end{pmatrix}; \quad \Sigma(\eta, s, e, i) = \begin{pmatrix} \lambda si & -\lambda si & 0 \\ -\lambda si & \lambda si + \epsilon e & -\epsilon e \\ 0 & -\epsilon e & \epsilon e + \gamma i \end{pmatrix}, \tag{6.2}$$

and the Cholesky decomposition of $\Sigma(\eta, \cdot)$ yields

$$\sigma(\eta, s, e, i) = \begin{pmatrix} \sqrt{\lambda si} & 0 & 0 \\ -\sqrt{\lambda si} & \sqrt{\epsilon e} & 0 \\ 0 & -\sqrt{\epsilon e} & \sqrt{\gamma i} \end{pmatrix}.$$

B.2 Details on the Kalman filter equations for incidence data of epidemic dynamics

Consider the model (3.17). Assume that $\mathcal{L}(\Delta_1 X) = \mathcal{N}_d(G_1, T_1)$ and $\mathcal{L}(Y_1|\Delta_1 X) = \mathcal{N}_q(B\Delta_1 X, P_1)$. Let $\widehat{\Delta_1 X} = G_1 = x(t_1) - x_0$ and $\widehat{\Xi}_1 = T_1$. Then, at iteration $k = 1$, the three steps of the Kalman filter are :

1. Prediction : $\mathcal{L}(\Delta_2 X|Y_1) = \mathcal{N}_d(\widehat{\Delta_2 X}, \widehat{\Xi}_2)$

$$\widehat{\Delta_2 X} = G_2 + (A_1 - I_d)\overline{\Delta_1 X}$$
$$\widehat{\Xi}_2 = (A_1 - I_d)\overline{T}_1(A_1 - I_d)' + T_2$$

2. Updating : $\mathcal{L}(\Delta_1 X|Y_1) = \mathcal{N}_d(\overline{\Delta_1 X}, \overline{T}_1)$

$$\overline{\Delta_1 X} = \widehat{\Delta_1 X} + \widehat{\Xi}_1 \tilde{B}'(\tilde{B}\widehat{\Xi}_1\tilde{B}' + \tilde{P}_1)^{-1}(Y_1 - \tilde{B}\widehat{\Delta_1 X})$$
$$\overline{T}_1 = \widehat{\Xi}_1 - \widehat{\Xi}_1\tilde{B}'(\tilde{B}\widehat{\Xi}_1\tilde{B}' + \tilde{P}_1)^{-1}\tilde{B}\widehat{\Xi}_1$$

3. Marginal : $\mathcal{L}(Y_2|Y_1) = \mathcal{N}(\widehat{M}_2, \widehat{\Omega}_2)$

$$\widehat{M}_2 = \tilde{B}\widehat{\Delta_2 X}$$
$$\widehat{\Omega}_2 = \tilde{B}\widehat{\Xi}_2\tilde{B}' + \tilde{P}_2$$

Now, starting from the distribution of $\mathcal{L}(\Delta_2 X|Y_1)$, the Kalman filter at iteration $k = 2$ becomes :

1. Prediction : $\mathcal{L}(\Delta_3 X|Y_2, Y_1) = \mathcal{N}_d(\widehat{\Delta_3 X}, \widehat{\Xi}_3)$

$$\widehat{\Delta_3 X} = G_3 + (A_2 - I_d)(\overline{\Delta_1 X} + \overline{\Delta_2 X})$$
$$\widehat{\Xi}_3 = (A_2 - I_d)(\overline{T}_1 + \overline{T}_2)(A_2 - I_d)' + T_3$$

2. Updating : $\mathcal{L}(\Delta_2 X|Y_2, Y_1) = \mathcal{N}_d(\overline{\Delta_2 X}, \overline{T}_2)$

$$\overline{\Delta_2 X} = \widehat{\Delta_2 X} + \widehat{\Xi}_2 \tilde{B}'(\tilde{B}\widehat{\Xi}_2\tilde{B}' + \tilde{P}_2)^{-1}(Y_2 - \tilde{B}\widehat{\Delta_2 X})$$
$$\overline{T}_2 = \widehat{\Xi}_2 - \widehat{\Xi}_2\tilde{B}'(\tilde{B}\widehat{\Xi}_2\tilde{B}' + \tilde{P}_2)^{-1}\tilde{B}\widehat{\Xi}_2$$

3. Marginal : $\mathcal{L}(Y_3|Y_2, Y_1) = \mathcal{N}(\widehat{M}_3, \widehat{\Omega}_3)$

$$\widehat{M}_3 = \tilde{B}\widehat{\Delta_3 X}$$
$$\widehat{\Omega}_3 = \tilde{B}\widehat{\Xi}_3\tilde{B}' + \tilde{P}_3$$

Proof : We just have to prove that, conditionally on Y_1, Y_2, $\Delta_1 X$ and $\Delta_2 X$ are independent. First, we have :

$$\Delta_3 X = G_3 + A_2(\Delta_1 X + \Delta_2 X) + U_3.$$

Hence :

$$\mathbb{E}(\Delta_3 X|Y_2, Y_1) = G_3 + A_2(\mathbb{E}(\Delta_1 X|Y_1) + \mathbb{E}(\Delta_2 X|Y_2, Y_1)) = G_3 + A_2(\overline{\Delta_1 X} + \overline{\Delta_2 X}).$$

Let $t_1, t_2 \in \mathbb{R}^d$. Then, we can compute the characteristic function of $\Delta_1 X + \Delta_2 X$ conditionally to Y_2, Y_1 :

$$\begin{aligned}\mathbb{E}\left[\exp\left(it_1'\Delta_1 X + it_2'\Delta_2 X\right)|Y_2, Y_1\right] &= \mathbb{E}\left[\exp\left(it_1'\Delta_1 X\right)|Y_2, Y_1\right]\mathbb{E}\left[\exp\left(it_2'\Delta_2 X|\Delta_1 X\right), Y_2, Y_1\right] \\ &= \exp\left(t_1'\overline{\Delta_1 X} + \frac{1}{2}t_1'\overline{T}_1\right) \times \exp\left(t_2'\overline{\Delta_2 X} + \frac{1}{2}t_2'\overline{T}_2\right).\end{aligned}$$

Consequently, conditionally to Y_1, Y_2, $\Delta_1 X$ and $\Delta_2 X$ are independent and

$$\mathrm{Var}(\Delta_1 X + \Delta_2 X|Y_2, Y_1) = \overline{T}_1 + \overline{T}_2.$$

□

Then, the generalization to the case $k \geq 1$ is direct, leading to the Kalman filter described in Section 3.3 for incidence data.

B.3 Practical considerations on implementation setting

Let us make some remarks on practical implementation.

— Two strategies for the choice of the step-size α_m at a given iteration m of the SAEM algorithm are combined, as recommended in (Lavielle (2014)) : first, denoting by M_0 the number of burn-in iterations, we use $\alpha_m = 1$ if $m \leq M_0$ to quickly converge to a neighborhood of the solution and then, $\alpha_m = \frac{1}{(m-M_0)^{\nu_0}}$ if $m > M_0$ with $\frac{1}{2} \leq \nu_0 \leq 1$ to ensure almost sure convergence of the sequence (θ_m) to the maximum likelihood estimate of θ.

— An extended algorithm for non-exponential models is proposed to include fixed effects (see e.g. Debavelaere and Allassonnière (2021)). Let κ be a fixed parameter to be estimated. First, for $m = 1, \ldots, M_0$, we use the classical procedure of the SAEM algorithm, that is a mean and a variance of the parameter is estimated at each iteration as if it were a random parameter. Then, at each new iteration $m+1$, the current variance of the parameter, denoted $\omega_\kappa^{(m+1)}$, is updated as : $\omega_\kappa^{(m+1)} = K_0 \times \omega_\kappa^{(m)}$, with $0 < K_0 < 1$.

— Due to the small influence of the number of iterations in the Metropolis-Hastings procedure (see e.g. Kuhn and Lavielle (2005)), a single iteration is used. Furthermore, if the proposal distribution is the marginal distribution $p(\mathbf{\Phi}; \tilde{\theta})$, the expression of the acceptance probability is simplified as follows :

$$\rho(\mathbf{\Phi}_{m-1}, \mathbf{\Phi}^{(c)}) = \min\left[1, \frac{p(\mathbf{y}|\mathbf{\Phi}^{(c)}; \tilde{\theta})}{p(\mathbf{y}|\mathbf{\Phi}_{m-1}; \tilde{\theta})}\right].$$

— A stopping criterion for the SAEM algorithm is considered. Denote by $\theta_j^{(m)}$ the j-th component of θ estimated at iteration m of the SAEM algorithm. Then, the algorithm stops either when the criterion

$$\max_j \left(\frac{|\theta_j^{(m)} - \theta_j^{(m-1)}|}{|\theta_j^{(m)}|} \right) < \mu_0$$

is satisfied several times consecutively or when a limit of $M_{\max}$ iterations is reached. The value of μ_0 is chosen sufficiently small (e.g. of the order of 10^{-3} or 10^{-4}).

— As the convergence of the SAEM algorithm can strongly depend on the initial guess, a simulated annealing version of SAEM (Kirkpatrick (1984)) is used to escape from potential local maxima of the likelihood during the first iterations and converge to a neighborhood of the global maximum. Let $\hat{\Gamma}\left(\phi_m^{(j)}\right)$ the estimated variance of the j-th component of $\mathbf{\Phi_m}$ at iteration m of the SAEM algorithm. Then, while $m \leq M_0$, $\Gamma_m^{(j)} = \max\left[\tau_0\, \Gamma_{m-1}^{(j)}, \hat{\Gamma}\left(\phi_m^{(j)}\right)\right]$ with $0 < \tau_0 < 1$. For $m > M_0$, the usual SAEM algorithm is used to estimate the variances at each iteration (see e.g. Lavielle (2014)).

— For the initialization of the SAEM algorithm, the starting parameter values β_0 of the fixed effects β are uniformly drawn from a hypercube encompassing the likely true values. The initial variances Γ_0 are chosen sufficiently large (1 by default).

— When the sampling intervals between observations Δ are large, the approximation of the resolvent matrix proposed in (Narci et al. (2021a)), Appendix A.1, is used.

— Concerning the KM approach, we use the Nelder-Mead method implemented in the `optim` function of the `R` software to maximize the approximated log-likelihood given by the Kalman filter. This requires to provide some initial values for the unknown parameters. As the optimization can be very sensitive to initialisation, 10 different starting values are considered and the maximum value for the log-likelihood among them are chosen. The starting parameter values for the maximization algorithm are uniformly drawn from a hypercube encompassing the likely true values.

For simulation studies in Section 3.4.1, the tuning parameters values are chosen as : $M_0 = 500$, $\nu_0 = 0.6$, $K_0 = 0.87$, $\mu_0 = 0.001$, $M_{\max} = 1000$ and $\tau_0 = 0.98$. Concerning the investigation of influenza outbreaks in Section 3.5, we chose : $M_0 = 5000$, $\nu_0 = 0.6$, $K_0 = 0.87$, $\mu_0 = 0.0001$ and $\tau_0 = 0.98$. The algorithm stops when the criterion is checked 100 times successively.

B.4 Estimation results for a second set of parameter values

B.4.1 Simulation settings

We consider a second set of parameter values which induces a lower intrinsic variability between epidemics. As for the first set of values, we consider two settings (denoted respectively (i) and (ii)) corresponding to two levels of inter-epidemic variability (resp. high and moderate) :

— Setting (i) : $\beta = (0.58, 1.10, 1.45, -2.20)'$ and $\Gamma = \text{diag}(0.47^2, 1.5^2, 0.75^2)$ corresponding to $\mathbb{E}(R_{0,1:U}) = 3$, $CV_{R_0} = 33\%$; $d = 3$; $\mathbb{E}(p_{1:U}) \approx 0.74$, $CV_p \approx 31\%$; $\mathbb{E}(i_{0,1:U}) \approx 0.12$, $CV_{i_0} \approx 66\%$.

— Setting (ii) : $\beta = (0.66, 1.10, 1.45, -2.2)'$ and $\Gamma = \text{diag}(0.25^2, 0.9^2, 0.5^2)$ corresponding to $\mathbb{E}(R_{0,1:U}) = 3$, $CV_{R_0} = 17\%$; $d = 3$; $\mathbb{E}(p_{1:U}) \approx 0.78$, $CV_p \approx 18\%$; $\mathbb{E}(i_{0,1:U}) \approx 0.11$, $CV_{i_0} \approx 45\%$.

B.4.2 Point estimates and standard deviation for inferred parameters

Tables 6.6 and 6.7 show the estimates of the expectation and standard deviation of the random effects ϕ_u, computed from the estimations of β and Γ using functions h defined in (3.24), for settings (i) and (ii). For each parameter, the reported values are the mean of the $J = 100$ parameter estimates $\phi_{u,j}$, $j \in \{1, \ldots, J\}$, and their standard deviations in brackets.

TABLE 6.6 – Estimates for setting (i) : high inter-epidemic variability. For each combination of $(\overline{n}, U)$ and for each model parameter (defined in the first line of the table), point estimates and precision are calculated as the mean of the $J = 100$ individual estimates and their standard deviations (in brackets).

Parameters		$\mathbb{E}(R_{0,u})$	d	$\mathbb{E}(p_u)$	$\mathbb{E}(i_{0,u})$	$\text{sd}(R_{0,u})$	$\text{sd}(p_u)$	$\text{sd}(i_{0,u})$
True values		**3.000**	**3.000**	**0.739**	**0.119**	**1.000**	**0.226**	**0.079**
$\overline{n} = 20$	$U = 20$	3.085	2.889	0.758	0.111	1.477	0.205	0.075
		(*0.460*)	(*0.205*)	(*0.060*)	(*0.016*)	(*0.666*)	(*0.036*)	(*0.018*)
	$U = 50$	3.152	2.926	0.761	0.111	1.509	0.199	0.075
		(*0.360*)	(*0.170*)	(*0.049*)	(*0.011*)	(*0.457*)	(*0.025*)	(*0.012*)
	$U = 100$	3.116	2.904	0.765	0.111	1.517	0.200	0.077
		(*0.307*)	(*0.152*)	(*0.046*)	(*0.008*)	(*0.366*)	(*0.018*)	(*0.009*)
$\overline{n} = 100$	$U = 20$	2.929	2.932	0.742	0.116	1.124	0.212	0.075
		(*0.263*)	(*0.144*)	(*0.047*)	(*0.016*)	(*0.332*)	(*0.029*)	(*0.017*)
	$U = 50$	3.002	2.973	0.749	0.116	1.186	0.207	0.075
		(*0.242*)	(*0.116*)	(*0.031*)	(*0.012*)	(*0.315*)	(*0.022*)	(*0.011*)
	$U = 100$	2.952	2.942	0.751	0.115	1.159	0.212	0.075
		(*0.148*)	(*0.090*)	(*0.022*)	(*0.008*)	(*0.155*)	(*0.018*)	(*0.007*)

TABLE 6.7 – Estimates for setting (ii) : moderate inter-epidemic variability. For each combination of $(\overline{n}, U)$ and for each model parameter (defined in the first line of the table), point estimates and precision are calculated as the mean of the $J = 100$ individual estimates and their standard deviations (in brackets).

Parameters		$\mathbb{E}(R_{0,u})$	d	$\mathbb{E}(p_u)$	$\mathbb{E}(i_{0,u})$	$\mathrm{sd}(R_{0,u})$	$\mathrm{sd}(p_u)$	$\mathrm{sd}(i_{0,u})$
True values		**3.000**	**3.000**	**0.777**	**0.109**	**0.500**	**0.143**	**0.049**
$\overline{n} = 20$	$U = 20$	3.183	3.051	0.771	0.106	0.811	0.128	0.046
		(*0.292*)	(*0.164*)	(*0.046*)	(*0.012*)	(*0.321*)	(*0.029*)	(*0.011*)
	$U = 50$	3.201	3.050	0.765	0.106	0.874	0.132	0.048
		(*0.208*)	(*0.116*)	(*0.035*)	(*0.008*)	(*0.241*)	(*0.018*)	(*0.007*)
	$U = 100$	3.232	3.068	0.765	0.106	0.906	0.132	0.048
		(*0.189*)	(*0.103*)	(*0.028*)	(*0.005*)	(*0.212*)	(*0.013*)	(*0.005*)
$\overline{n} = 100$	$U = 20$	3.037	3.051	0.770	0.110	0.563	0.135	0.046
		(*0.169*)	(*0.100*)	(*0.037*)	(*0.012*)	(*0.206*)	(*0.026*)	(*0.011*)
	$U = 50$	3.064	3.055	0.764	0.110	0.632	0.139	0.048
		(*0.117*)	(*0.080*)	(*0.023*)	(*0.009*)	(*0.142*)	(*0.016*)	(*0.007*)
	$U = 100$	3.059	3.057	0.768	0.110	0.619	0.141	0.048
		(*0.088*)	(*0.061*)	(*0.019*)	(*0.005*)	(*0.094*)	(*0.013*)	(*0.004*)

As for the first set of parameters values, all point estimates are closed to the true values. The standard error of the estimates decreases when the number of epidemics U and the number of observations $\bar{n}$ increases, whereas the bias is only sensitive to $\bar{n}$ (bias decreasing when $\bar{n}$ increasing).

For a given data set, Figure 6.4 displays convergence graphs for model parameters in setting (i) with $U = 100$ and $\bar{n} = 100$.

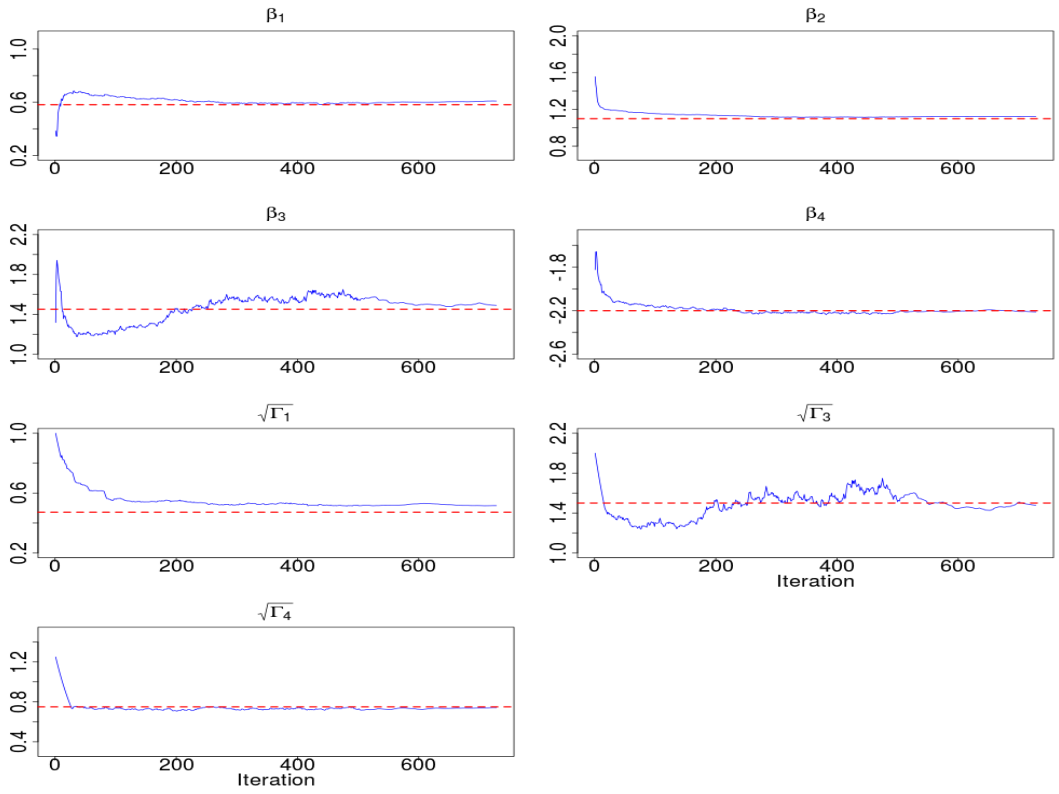

FIGURE 6.4 – Convergence graphs of the SAEM algorithm for estimates of $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$ and $\text{diag}(\Gamma) = (\Gamma_1, \Gamma_3, \Gamma_4)$. Setting (i) with $U = 100$ and $\bar{n} = 100$. Parameter values at each iteration of the SAEM algorithm (plain blue line) and true values of model parameters (dotted red line).

We notice that all model parameters converge towards their true value.

C Annexes : Chapitre 4

C.1 Modèles mécanistes utilisés dans la littérature pour décrire la propagation du SARS-CoV-2 dans une population en France

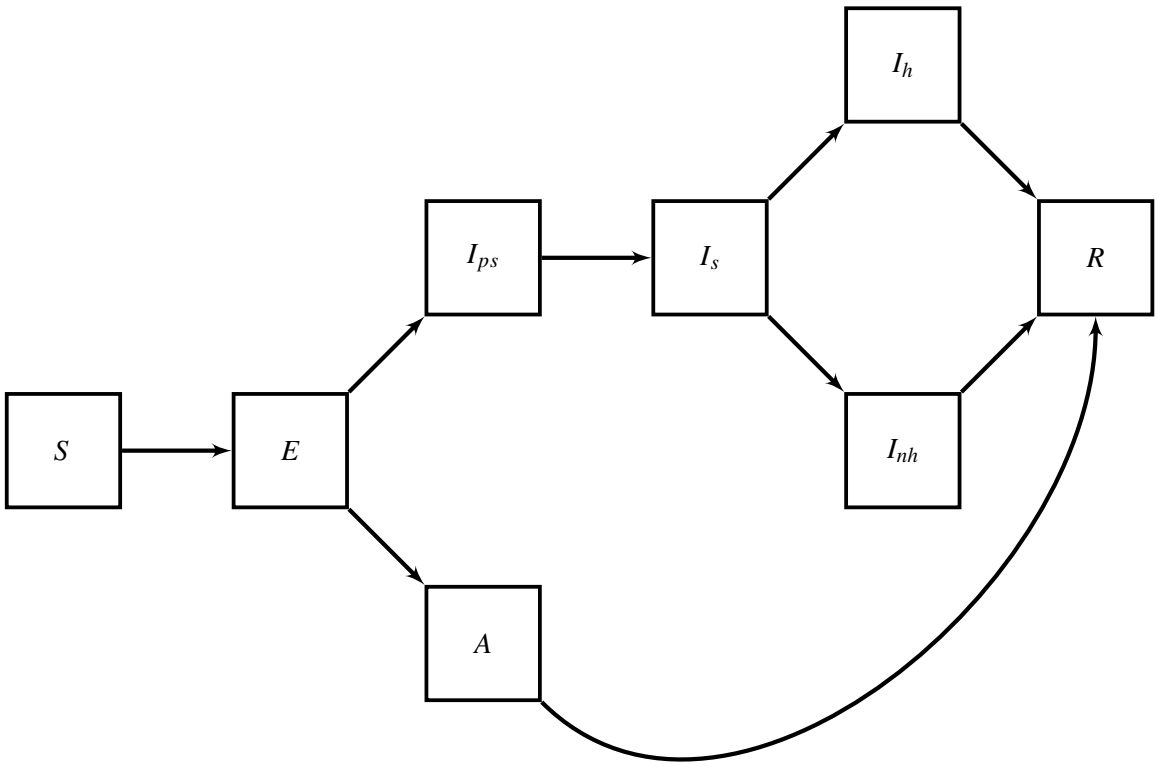

FIGURE 6.5 – Diagramme du modèle épidémiologique utilisé par Roux et al. (2021). Les individus peuvent être susceptibles (S), exposés (E), infectés et pré-symptomatiques (I_{ps}), infectés et asymptomatiques (A), infectés symptomatiques (I_s), hospitalisés (I_h), non-hospitalisés (I_{nh}) ou retirés de la chaîne de transmission (R), *i.e.* guéris ou décédés.

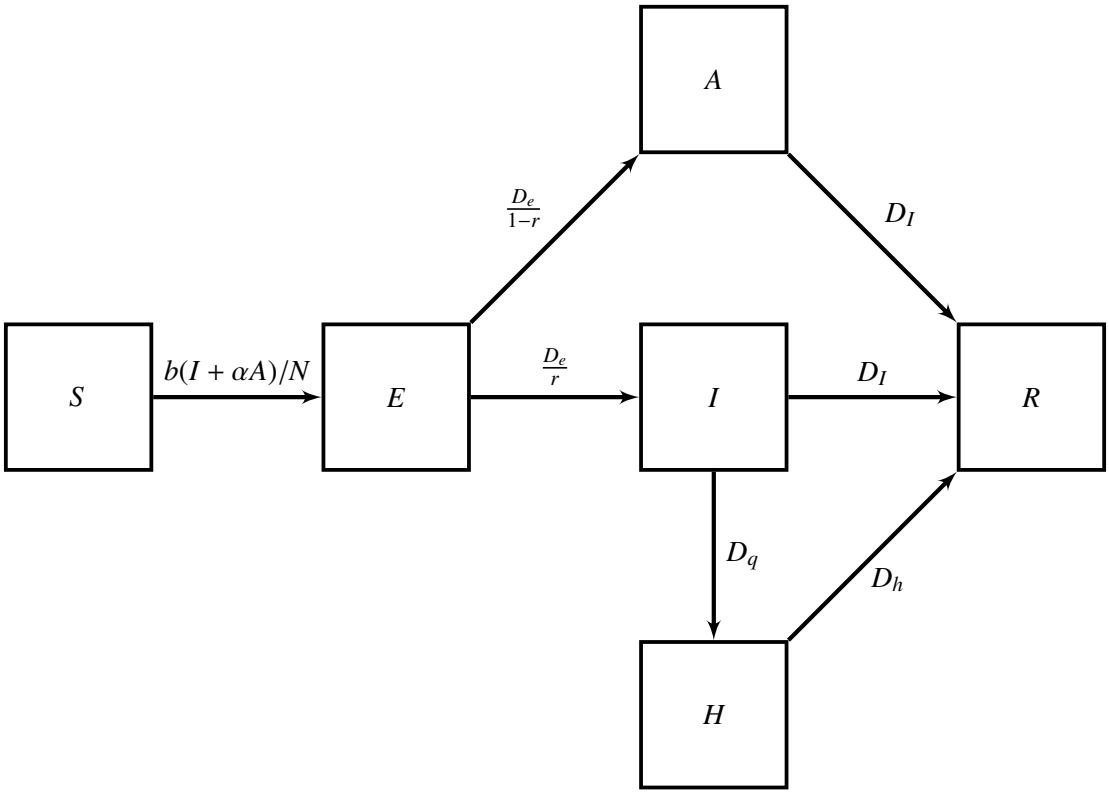

FIGURE 6.6 – Modèle à compartiments SEIRAH utilisé par Prague et al. (2020), où les compartiments I et A correspondent respectivement à des individus dont l'infection est avérée ou non. Autrement dit, un individu appartenant au compartiment A peut être soit asymptomatique et non-testé, soit symptomatique et non-testé. S, E, H, et R correspondent respectivement à des individus susceptibles, exposés, hospitalisés et retirés de la chaîne de transmission (*i.e.* guéris ou décédés). En reprenant les notations des auteurs, les paramètres du modèle sont : b le taux de transmission, α la réduction de transmission chez les individus asymptomatiques, D_e la durée d'exposition, D_I la durée d'infectiosité, D_q la durée entre le début des symptômes et l'hospitalisation, D_h la durée d'hospitalisation et r le taux auquel un cas est documenté.
Dans (Collin et al. (2021)), le modèle est similaire mais avec quelques nuances. Tout d'abord, les compartiments A et I correspondent respectivement à des individus infectieux asymptomatiques et infectieux symptomatiques. De plus, le taux de transition entre S et E, prenant en compte le nombre d'individus vaccinés V considéré connu, devient $b\left(1-\frac{V}{N}\right)\frac{I+\alpha A}{N}$. Enfin, un individu exposé a une probabilité $(1-r_E)$ d'être asymptomatique et un individu infectieux a une probabilité $(1-r_I)$ d'être hospitalisé.

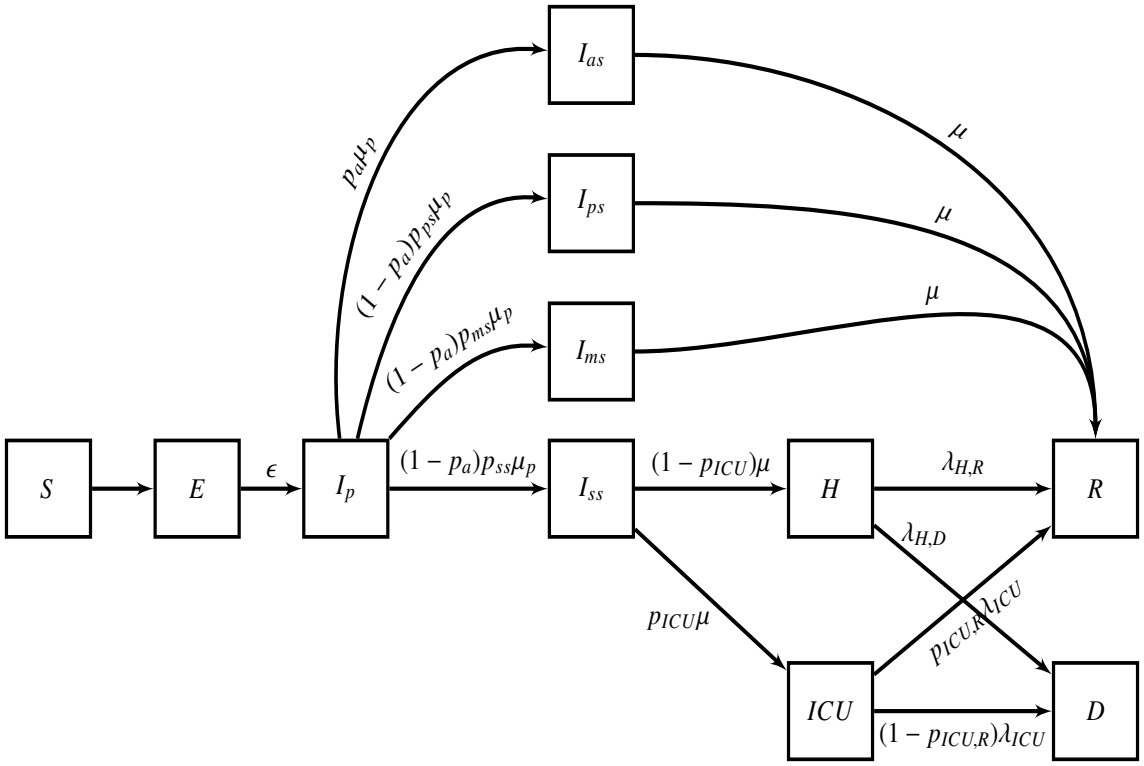

FIGURE 6.7 – Modèle à compartiments utilisé par Di Domenico et al. (2020) où les individus peuvent être susceptibles (S), exposés (E), infectieux pré-symptomatiques (I_p), infectieux asymptomatiques (I_{as}), infectieux paucisymptomatiques (I_{ps}), infectieux symptomatiques avec des symptômes légers (I_{ms}) ou sévères (I_{ss}), hospitalisés (H), en unité de soins intensifs (ICU), décédés (D) ou guéris (R). Les paramètres du modèle sont : ϵ le taux d'exposition, p_a la probabilité d'être asymptomatique (resp. p_{ps}, p_{ms}, p_{ss}, p_{ICU} et $p_{ICU,R}$ la probabilité d'être paucisymptomatique, faiblement symptomatique, sévèrement symptomatique, d'aller en unité de soins intensifs et de guérir après être passé en unité de soins intensifs) et μ_p^{-1} (resp. μ^{-1}, $\lambda_{H,R}^{-1}$ et $\lambda_{H,D}^{-1}$) la durée entre la phase pré-symptomatique et la phase symptomatique ou asymptomatique (resp. la durée de guérison pour les individus infectieux sans symptômes sévères ou la durée avant hospitalisation pour ceux avec symptômes sévères, la durée entre H et R et la durée entre H et D).

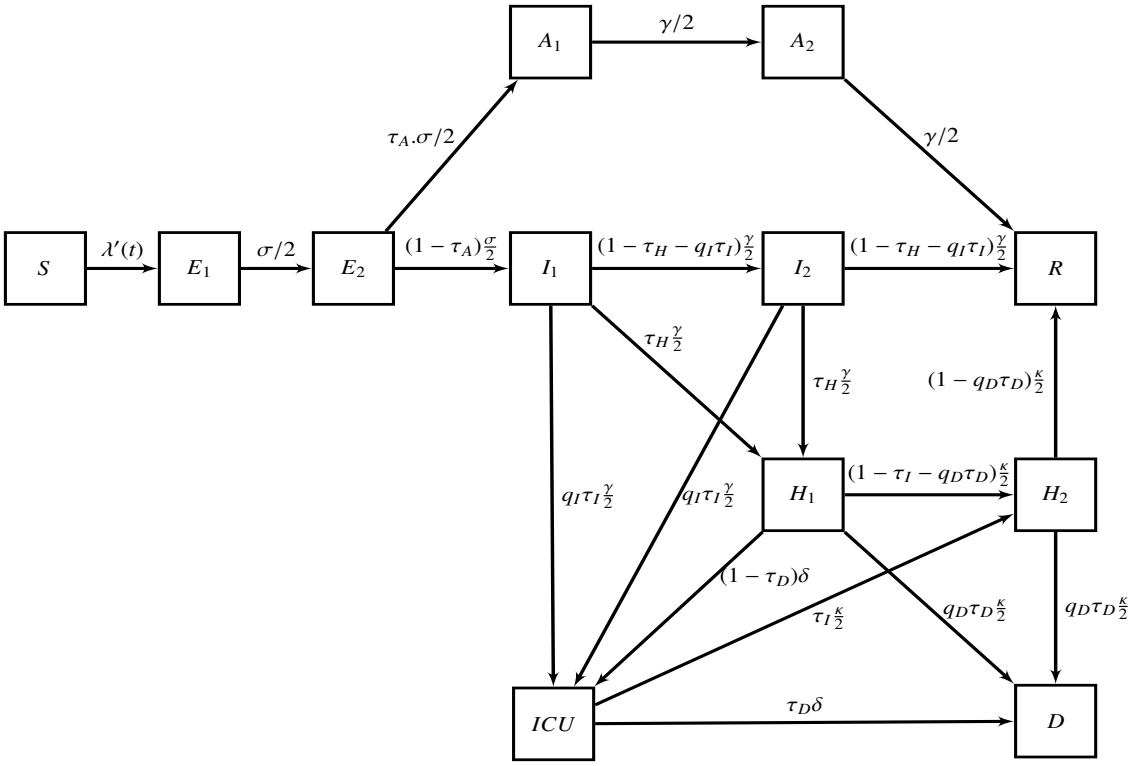

FIGURE 6.8 – Modèle à compartiments utilisé par Cazelles et al. (2021) où les individus peuvent être susceptibles (S), exposés (E), infectieux asymptomatiques (A), infectieux symptomatiques (I), hospitalisés (H), en unité de soins intensifs (ICU), décédés (D) ou guéris (R). Les indices 1 et 2, correspondant aux sous-compartiments, pour E, A, I et H sont utilisés afin que les temps de séjour dans ces compartiments suivent une loi d'Erlang. Les paramètres du modèle sont : $\lambda(t) = \lambda'(t)S(t)$ la force d'infection avec $\lambda'(t) = \beta(t)\frac{I_1+q_1I_2+q_2(A_1+A_2)}{N}$, $\beta(t)$ le taux de transmission modélisé par un processus de diffusion, σ le taux d'exposition, γ le taux de guérison, κ^{-1} la durée d'hospitalisation, δ^{-1} la durée passée en ICU, τ_A, τ_H, τ_I et τ_D la fraction d'individus asymptomatiques, hospitalisés, admis en ICU et décédés, q_1 et q_2 la réduction dans la transmissibilité de I_2 et $(A_i)_{i=1,2}$, q_I la réduction dans la fraction d'individus admis en ICU et q_D la réduction dans le taux de mortalité.

C.2 Tableau résumé de plusieurs études d'inférence

Le Tableau 6.8 réunit plusieurs articles, publiés ou sous format preprint, avec application d'une méthode d'inférence sur des données Covid-19 en France, pour lesquels sont indiqués le modèle et les données utilisées ainsi que les paramètres connus ou inconnus et estimés. Comme les modèles présentés peuvent être sensiblement différents, les paramètres mécanistes impliqués n'ont pas toujours la même signification. Quand le lien entre ceux-ci et les paramètres décrits dans la Figure 4.2 est possible à discerner, nous indiquons les valeurs considérées connues et celles estimées de ces paramètres.

TABLE 6.8 – Paramètres mécanistes, avec valeurs connues et fixées ou inconnues et estimées, de plusieurs articles mettant en oeuvre une méthode d'inférence sur des données Covid-19 en France. Les données d'hôpital correspondent à toutes celles en lien avec l'hôpital (e.g. le nombre de patients hospitalisés et/ou en réanimation, le nombre de décès). Selon l'article, la méthode d'inference est appliquée sur un nombre $U \geq 1$ de régions (IDF = Île-de-France, PACA = Provence-Alpes-Côte d'Azur, OC = Occitanie, NA = Nouvelle-Aquitaine, ARA = Auvergne-Rhône-Alpes). Pour les articles traitant $U > 1$ régions et pour un paramètre générique noté $\psi = (\psi_1, \ldots, \psi_U)$, la moyenne régionale du paramètre ψ est noté $\bar{\psi} = \sum_{u=1}^{U} \psi_u$ et est calculée empiriquement dans le cas où aucune distribution ne lui est spécifiée (écart-type empirique entre parenthèses). Dans le cadre des modèles à effets mixtes, le paramètre ψ étant une variable aléatoire, les valeurs fournies correspondent à son espérance $\mathbb{E}(\psi_u)$ et son écart-type $\mathrm{sd}(\psi_u)$.

Référence	Modèle	Données	Paramètres connus	Paramètres estimés
Roux et al. (2021)	cf Figure 6.5	Hôpital $U = 13$ régions Printemps 2020	$d_E = 2.72$, $d_I = 10.91$ $\alpha = 0.55$	$\bar{R}_0 = 2.60$ $1.94 \leq R_{0,u} \leq 4.17\ \forall u$
Prague et al. (2020)	cf Figure 6.6	Hospitalisations, infections $U = 12$ régions 02/03/2020-11/05/2020	$\alpha = 0.55$, $d_E = 5.1$, $d_I = 2.3$, $d_H = 30$	$0.17 \leq \tau_{A,u} \leq 0.19\ \forall u$ $\mathbb{E}(R_{0,u}) = 2.81$ $\mathbb{E}(\lambda_u) = 2.23$, $\mathrm{sd}(\lambda_u) = 0.003$
Di Domenico et al. (2020)	cf Figure 6.7	Hôpital $U = 1$ région : IDF 01/03/2020-23/03/2020	$\tau_A \in \{0.2, 0.5\}$, $d_E = 3.7$, $d_I = 3.8$	$R_0 = 3.18$
Collin et al. (2021)	cf Figure 6.6	Hospitalisations $U = 12$ régions 02/03/2020-28/03/2021	$\tau_A = 0.156$, $\tau_H = 0.034$, $\alpha = 0.55$, $d_E = 5.1$ $d_I = 5$, $d_H = 18.3$	$3.5 \leq R_{0,u} \leq 4\ \forall u$ $\mathbb{E}(\lambda_u) = 0.78$, $\mathrm{sd}(\lambda_u) = 0.012$
Cazelles et al. (2021)	cf Figure 6.8	Hôpital, infections $U = 5$ régions : IDF, PACA, OC, NA, ARA 28/02/2020-début 2021	$\bar{\lambda} = 0.76$ (0.08) $\alpha = 0.55$	$\bar{d}_{E,U} = 4$ (0.008) $\bar{d}_{I,U} = 5.98$ (0.04) $\bar{d}_{H,U} = 13.35$ (1.50) $\bar{\tau}_{A,U} = 0.52$ (0.03) $\bar{\tau}_{H,U} = 0.025$ (0.003) $\bar{\tau}_{D,U} = 0.042$ (0.005)

C.3 Visualisation de l'évaluation post-prédictive pour $\tau_A = 0.4$ et $\tau_A = 0.6$

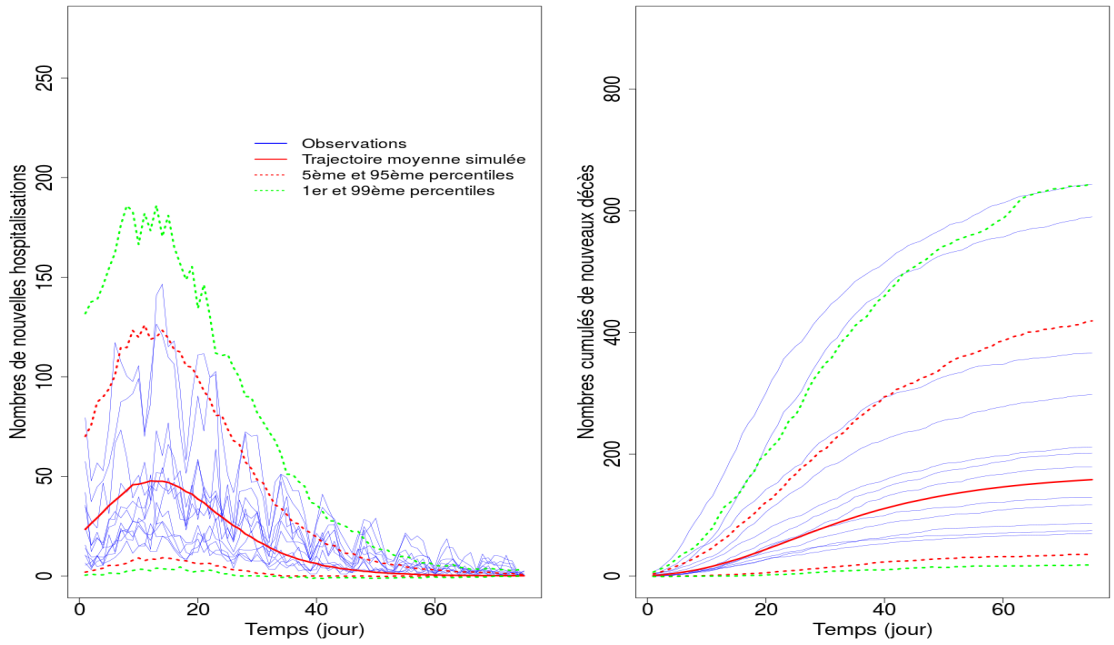

FIGURE 6.9 – Visualisation de l'évaluation post-prédiction avec $\tau_A = 0.4$ et $d_H = 20$ jours. Observations : nombres quotidiens de nouvelles hospitalisations et nombres cumulés de nouveaux décès calculés sur $1,000,000$ d'habitants pour chacune des U épidémies (bleu). Trajectoires simulées obtenues en trois étapes : (i) génération de 1000 valeurs de $\hat{\phi}_u$ en fonction des valeurs estimées des paramètres du modèle ; (ii) sachant $\hat{\phi}_u$, simulation de 1000 épidémies (processus Markoviens de sauts) selon le modèle épidémique ; (iii) calcul de la trajectoire moyenne (ligne pleine rouge), des 5ème et 95ème percentiles (pointillé rouge) et des 1er et 99ème percentiles (pointillé vert) sur les 1000 épidémies simulées.

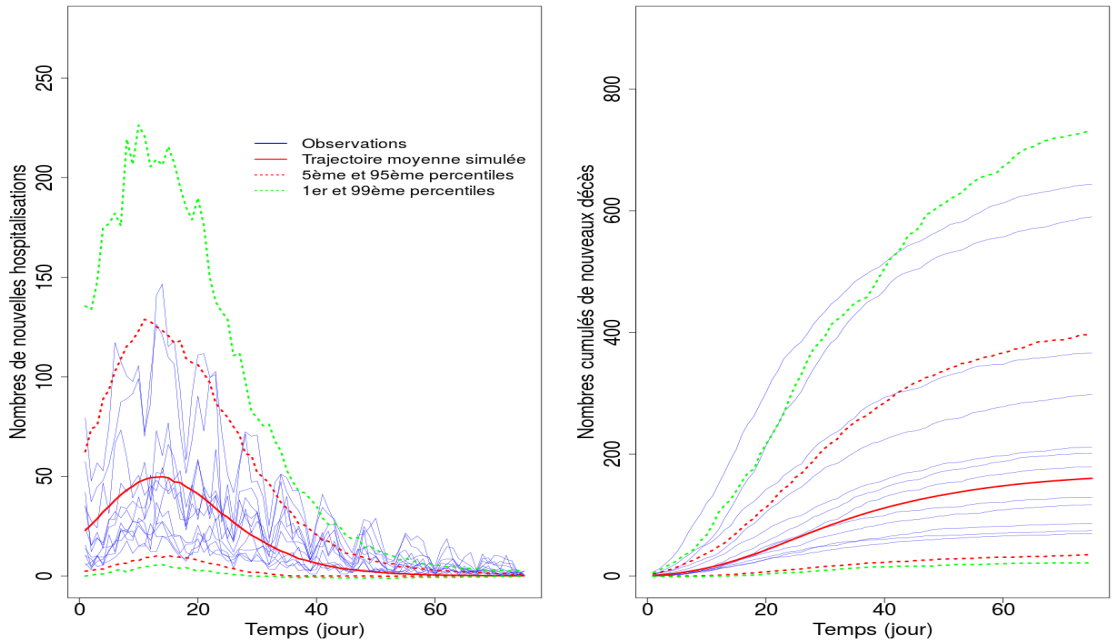

FIGURE 6.10 – Visualisation de l'évaluation post-prédiction avec $\tau_A = 0.6$ et $d_H = 20$ jours. Observations : nombres quotidiens de nouvelles hospitalisations et nombres cumulés de nouveaux décès calculés sur $1,000,000$ d'habitants pour chacune des U épidémies (bleu). Trajectoires simulées obtenues en trois étapes : (i) génération de 1000 valeurs de $\hat{\phi}_u$ en fonction des valeurs estimées des paramètres du modèle ; (ii) sachant $\hat{\phi}_u$, simulation de 1000 épidémies (processus Markoviens de sauts) selon le modèle épidémique ; (iii) calcul de la trajectoire moyenne (ligne pleine rouge), des 5ème et 95ème percentiles (pointillé rouge) et des 1er et 99ème percentiles (pointillé vert) sur les 1000 épidémies simulées.

Bibliographie

R. M. Anderson, R. M. May, T. W. Ng, and J. T. Rowley. Age-dependent choice of sexual partners and the transmission dynamics of hiv in sub-saharan africa. Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences, 336(1277) :135–155, 1992. doi : 10.1098/rstb.1992.0052.

H. Andersson and T. Britton. Stochastic epidemic models and their statistical analysis, volume 151 of Lecture Notes in Statistics. Springer, 2000. doi : 10.1007/978-1-4612-1158-7.

C. Andrieu, A. Doucet, and R. Holenstein. Particle markov chain monte carlo methods. Journal of the Royal Statistical Society B, 72(3) :269–342, 2010. doi : 10.1111/j.1467-9868.2009.00736.x.

Anonymous. Influenza in a boarding school. British Medical Journal, 1978.

R. Azencott. Formule de taylor stochastique et développement asymptotique intégrales de feynmann. Séminaire de Probabilités XVI, pages 237–285., 1982. URL http://www.numdam.org/item/SPS_1982__S16__237_0.

M. Baguelin, A. J. V. Hoek, M. Jit, S. Flasche, P. J. White, and W. J. Edmunds. Vaccination against pandemic influenza a/h1n1v in england : a real-time economic evaluation. Vaccine, 28 (12) :2370–2384, 2010. doi : 10.1016/j.vaccine.2010.01.002.

M. Baguelin, S. Flasche, A. Camacho, N. Demiris, E. Miller, and W. J. Edmunds. Assessing optimal target populations for influenza vaccination programmes : An evidence synthesis and modelling study. PLOS Medicine, 10(10), 2013. doi : 10.1371/journal.pmed.1001527.

N. T. J. Bailey. The Mathematical Theory of Infectious Diseases and its Applications. London : Charles Griffin and Company, 1975.

O. N. Bjørnstad, B. F. Finkenstädt, and B. T. Grenfell. Dynamics of measles epidemics : Estimating scaling of transmission rates using a time series sir model. Ecological Monographs, 72(2) : 169–184, 2002. doi : 10.2307/3100023.

J. C. Blackwood, D. A. T. Cummings, H. Broutin, S. Iamsirithaworn, and P. Rohani. Deciphering the impacts of vaccination and immunity on pertussis epidemiology in thailand. Proceedings of the National Academy of Sciences of the United States of America, 110(23) :9595–9600, 2013a. doi : 10.1073/pnas.1220908110.

J. C. Blackwood, D. G. Streicker, S. Altizer, and P. Rohani. Resolving the roles of immunity, pathogenesis, and immigration for rabies persistence in vampire bats. Proceedings of the National Academy of Sciences of the United States of America, 110(51) :20837–20842, 2013b. doi : 10.1073/pnas.1308817110.

I. M. Blake, R. Martin, A. Goel, N. Khetsuriani, J. Everts, C. Wolff, S. Wassilak, R. B. Aylward, and N. C. Grassly. The role of older children and adults in wild poliovirus transmission. Proceedings of the National Academy of Sciences of the United States of America, 111(29) : 10604–10609, 2014. doi : 10.1073/pnas.1323688111.

François Blanquart, Nathanaël Hozé, Benjamin J. Cowling, Florence Débarre, and Simon Cauchemez. Selection for infectivity profiles in slow and fast epidemics, and the rise of sars-cov-2 variants. medRxiv, 2021. doi : 10.1101/2021.12.08.21267454v1.

C. Breto, D. He, E. L. Ionides, and A. A. King. Time series analysis via mechanistic models. The Annals of Applied Statistics, 3(1) :319–348, 2009. doi : 10.1214/08-AOAS201.

C. Bretó. Modeling and inference for infectious disease dynamics : A likelihood-based approach. Stat. Sci., 33(1) :57–69, 2018. doi : 10.1214/17-STS636.

C. Bretó, E. L. Ionides, and A. A. King. Panel data analysis via mechanistic models. JASA, 115 (531) :1178–1188, 2020. doi : 10.1080/01621459.2019.1604367.

T. Britton. Stochastic epidemic models : A survey. Mathematical Biosciences, 225(1) :24–35, 2010. doi : https://doi.org/10.1016/j.mbs.2010.01.006.

T. Britton and F. Giardina. Introduction to statistical inference for infectious diseases. Journal de la Société Française de Statistique, 157(1) :53–70, 2016.

T. Britton and E. Pardoux. Stochastic epidemic models with inference. Springer, 2020. doi : 10.1007/978-3-030-30900-8.

E. Buckingham-Jeffery, V. Isham, and T. House. Gaussian process approximations for fast inference from infectious disease data. Mathematical Biosciences, 301 :111 – 120, 2018. ISSN 0025-5564. doi : 10.1016/j.mbs.2018.02.003.

A. Camacho, S. Ballesteros, A. L. Graham, F. Carrat, O. Ratmann, and B. Cazelles. Explaining rapid reinfections in multiple-wave influenza outbreaks : Tristan da cunha 1971 epidemic as a case study. Proceedings of the Royal Society B : Biological Sciences, 278(1725) :3635–3643, 2011. doi : 10.1098/rspb.2011.0300.

O. Cappé, E. Moulines, and T. Rydén. Inference in Hidden Markov Models. Springer, 2005.

F. Carrat, E. Vergu, N. M. Ferguson, M. Lemaitre, S. Cauchemez, S. Leach, and A-J Valleron. Time Lines of Infection and Disease in Human Influenza : A Review of Volunteer Challenge Studies. American Journal of Epidemiology, 167(7) :775–785, 2008. ISSN 0002-9262. doi : 10.1093/aje/kwm375.

D. Carslake, J. Cave, W. Grant, J. Greaves, L. Green, M. Keeling, J. McEldowney, G. Medley, and H. Weldegebriel. Animal health and welfare : A case study of science, law and policy in a regulatory environment. 3, 2010.

W. Cates, R. B. Rothenberg, and J. H. Blount. Syphilis control. the historic context and epidemiologic basis for interrupting sexual transmission of treponema pallidum. Sexually Transmitted Diseases, 23(1) :68–75, 1996. doi : 10.1097/00007435-199601000-00013.

S. Cauchemez and N. M. Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data : application to measles transmission in london. Journal of The Royal Society Interface, 5(25) :885–897, 2008. doi : 10.1098/rsif.2007.1292.

S. Cauchemez, A. J. Valleron, P. Y. Boëlle, A. Flahault, and N. M. Ferguson. Estimating the impact of school closure on influenza transmission from sentinel data. Nature, 452 :750–754, 2008. doi : 10.1038/nature06732.

B. Cazelles, M. Chavez, A. J. McMichael, and S. Hales. Nonstationary influence of el niño on the synchronous dengue epidemics in thailand. PLOS Medicine, 2(4) :e106, 2005. doi : 10.1371/journal.pmed.0020106.

B. Cazelles, C. Champagne, B. N. V. Yen, C. Comiskey, E. Vergu, and B. Roche. A mechanistic and data-driven reconstruction of the time-varying reproduction number : Application to the covid-19 epidemic. medRxiv, 2021. doi : 10.1101/2021.02.04.21251167.

K. Chan and J. Ledolter. Monte carlo em estimation for time series models involving counts. Journal of The American Statistical Association, 90 :242–252, 1995. doi : 10.1080/01621459.1995.10476508.

G. Chowell, M. A. Miller, and C. Viboud. Seasonal influenza in the united states, france, and australia : transmission and prospects for control. Epidemiology and Infection, 136(6) :852–864, 2008. doi : 10.1017/S0950268807009144.

A. Collin, M. Prague, and P. Moireau. Estimation for dynamical systems using a population-based kalman filter - applications to pharmacokinetics models. 2020. URL https://hal.inria.fr/hal-02869347. Working paper or preprint.

A. Collin, B. Hejblum, C. Vignals, L. Lehot, R. Thiebaut, P. Moireau, and M. Prague. Using population based kalman estimator to model covid-19 epidemics in france : estimating the burden of sars-cov-2 and the effects of npi. medRxiv, 2021. doi : 10.1101/2021.07.09.21260259.

A. Cori, A. J. Valleron, F. Carrat, G. Scalia-Tomba, G. Thomas, and P. Y. Boëlle. Estimating influenza latency and infectious period durations using viral excretion data. Epidemics, 4(3) : 132–138, 2012. ISSN 1755-4365. doi : 10.1016/j.epidem.2012.06.001.

V. Debavelaere and S. Allassonnière. On the curved exponential family in the Stochatic Approximation Expectation Maximization Algorithm. February 2021. URL https://hal.archives-ouvertes.fr/hal-03128554, preprint.

M. Delattre and M. Lavielle. Coupling the saem algorithm and the extended kalman filter for maximum likelihood estimation in mixed-effects diffusion models. Statistics and Its Interface, 6 :519–532, 2013. doi : 10.4310/SII.2013.v6.n4.a10.

M. Delattre and M-A Poursat. An iterative algorithm for joint covariate and random effect selection in mixed effects models. The International Journal of Biostatistics, 16(2) :1–12, 2020. doi : 10.1515/ijb-2019-0082.

M. Delattre, M. Lavielle, and M-A Poursat. A note on BIC in mixed-effects models. EJS, 8(1) : 456–475, 2014. doi : 10.1214/14-EJS890.

M. Delattre, V. Genon-Catalot, and C. Larédo. Parametric inference for discrete observations of diffusion processes with mixed effects. Stochastic Processes and their Applications, 128(6) : 1929–1957, 2018. ISSN 0304-4149. doi : 10.1016/j.spa.2017.08.016.

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. Ann. Statist., 27(1) :94–128, 1999. doi : 10.1214/aos/1018031103.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1) :1–38, 1977. ISSN 00359246. URL http://www.jstor.org/stable/2984875.

L. Di Domenico, G. Pullano, C. E. Sabbatini, P-Y Boëlle, and V. Colizza. Impact of lockdown on covid-19 epidemic in Île-de-france and possible exit strategies. BMC Medicine, 18(1) :240, 2020. doi : 10.1186/s12916-020-01698-4.

X. Didelot, L. K. Whittles, and I. Hall. Model-based analysis of an outbreak of bubonic plague in cairo in 1801. Journal of The Royal Society Interface, 14(131) :20170160, 2017. doi : 10.1098/rsif.2017.0160.

C. A. Donnelly, A. C. Ghani, G. M. Leung, A. J. Hedley, C. Fraser, S. Riley, L. J. Abu-Raddad, L-M Ho, T-Q Thach, P. Chau, and et al. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in hong kong. Lancet (London, England), 361(9371) : 1761–1766, 2003. doi : 10.1016/S0140-6736(03)13410-1.

S. Donnet and A. Samson. Estimation of parameters in incomplete data models defined by dynamical systems. Journal of Statistical Planning and Inference, 137(9) :2815–2831, 2007. doi : 10.1016/j.jspi.2006.10.013.

S. Donnet and A. Samson. Parametric inference for mixed models defined by stochastic differential equations. ESAIM : PS, 12 :196–218, 2008. doi : 10.1051/ps:2007045.

S. Donnet and A. Samson. A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models. Advanced Drug Delivery Reviews, 65(7) :929–939, 2013. ISSN 0169-409X. doi : 10.1016/j.addr.2013.03.005.

S. Donnet and A. Samson. Using pmcmc in em algorithm for stochastic mixed models : theoretical and practical issues. Journal de la Société Française de Statistique, 155(1) :49–72, 2014. URL http://www.numdam.org/item/JSFS_2014__155_1_49_0/.

A. Doucet, N. de Freitas, and N. Gordon. Sequential Monte Carlo Methods in Practice. Springer, 2001. doi : 10.1007/978-1-4757-3437-9.

S. N. Ethier and T. G. Kurtz. Markov processes : characterization and convergence. Wiley, 2nd edition., 2005. doi : 10.1002/9780470316658.

B. Favetto and A. Samson. Parameter estimation for a bidimensional partially observed ornstein-uhlenbeck process with biological application. Scandinavian Journal of Statistics, 37(2) :200–220, 2010. doi : 10.1111/j.1467-9469.2009.00679.x.

N. M. Ferguson, D. J. Nokes, and R. M. Anderson. Dynamical complexity in age-structured models of the transmission of the measles virus : Epidemiological implications at high levels of vaccine uptake. Mathematical Biosciences, 138(2) :101–130, 1996. doi : 10.1016/S0025-5564(96)00127-7.

N. M. Ferguson, C. A. Donnelly, M. E. J. Woolhouse, and R. M. Anderson. The epidemiology of bse in cattle herds in great britain. ii. model construction and analysis of transmission dynamics. Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences, 352 (1355) :803–838, 1997. doi : 10.1098/rstb.1997.0063.

N. M. Ferguson, M. J. Keeling, W. John Edmunds, R. Gani, B. T. Grenfell, R. M. Anderson, and S. Leach. Planning for smallpox outbreaks. Nature, 425(6959) :681–685, 2003. doi : 10.1038/nature02007.

N. M. Ferguson, A. T. D. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke. Strategies for containing an emerging influenza pandemic in southeast asia. Nature, 437 :209–214, 2005. doi : 10.1038/nature04017.

P. E. Fine and J. A. Clarkson. Measles in england and wales–i : An analysis of factors underlying seasonal patterns. International Journal of Epidemiology, 11(1) :5–14, 1982. doi : 10.1093/ije/11.1.5.

B. F. Finkenstädt and B. T. Grenfell. Time series modelling of childhood diseases : a dynamical systems approach. Journal of the Royal Statistical Society : Series C (Applied Statistics), 49 (2) :187–205, 2000. doi : 10.1111/1467-9876.00187.

G. Fort and E. Moulines. Convergence of the monte carlo expectation maximization for curved exponential families. The Annals of Statistics, 31(4) :1220–1259, 2003. doi : 10.1214/aos/1059655912.

C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, T. Jombart, W. R. Hinsley, N. C. Grassly, F. Balloux, A. C. Ghani, N. M. Ferguson, A. Rambaut, O. G. Pybus, H. Lopez-Gatell, C. M. Alpuche-Aranda, I. B. Chapela, E. P. Zavala, D. M. Guevara, F. Checchi, E. Garcia, S. Hugonnet, and C. Roth. Pandemic potential of a strain of influenza a (h1n1) : early findings. Science, 324(5934) :1557–1561, 2009. doi : 10.1126/science.1176062.

M. Freidlin and A. Wentzell. Random perturbations of dynamical systems. Springer, 1978. doi : 10.1007/978-3-642-25847-3.

C. Fuchs. Inference for diffusion processes. Springer, 2013. doi : 10.1007/978-3-642-25969-2.

S. Funk, A. Camacho, A. J. Kucharski, R. M. Eggo, and W. J. Edmunds. Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. Epidemics, 22 :56–61, 2018. doi : 10.1016/j.epidem.2016.11.003.

A. Gaymard, P. Bosetti, A. Feri, G. Destras, V. Enouf, A. Andronico, S. Burrel, S. Behillil, C. Sauvage, A. Bal, and et al. Early assessment of diffusion and possible expansion of sars-cov-2 lineage 20i/501y.v1 (b.1.1.7, variant of concern 202012/01) in france, january to march 2021. Eurosurveillance, 26(9) :2100133, 2021. doi : 10.2807/1560-7917.ES.2021.26.9.2100133.

I. Ghosh, P. K. Tiwari, S. Samanta, I. M. Elmojtaba, N. Al-Salti, and J. Chattopadhyay. A simple si-type model for hiv/aids with media and self-imposed psychological fear. Mathematical Biosciences, 306 :160–169, 2018. doi : 10.1016/j.mbs.2018.09.014.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. J. Chem. Phys, 81 (25) :2340–2361, 1977. doi : 10.1021/j100540a008.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. The Journal of Chemical Physics, 115(4) :1716–1733, 2001. doi : 10.1063/1.1378322.

R. Guy, C. Larédo, and E. Vergu. Parametric inference for discretely observed multidimensional diffusions with small diffusion coefficient. Stochastic Processes and their Applications, 124(1) : 51–80, 2014. doi : 10.1016/j.spa.2013.07.009.

R. Guy, C. Larédo, and E. Vergu. Approximation of epidemic models by diffusion processes and their statistical inference. J. Math. Bio, 70(3) :621–646, 2015. doi : 10.1007/s00285-014-0777-8.

D. He, E. L. Ionides, and A. A. King. Plug-and-play inference for disease dynamics : measles in large and small populations as a case study. Journal of The Royal Society Interface, 7(43) : 271–283, 2010. doi : 10.1098/rsif.2009.0151.

H. Heesterbeek, R. M. Anderson, V. Andreasen, S. Bansal, D. De Angelis, C. Dye, K. T. D. Eames, W. J. Edmunds, S. D. W. Frost, S. Funk, and et al. Modeling infectious disease dynamics in the complex landscape of global health. Science, 347(6227) :aaa4339, 2015. doi : 10.1126/science.aaa4339.

C. Hsiao. Analysis of Panel Data. Econometric Society Monographs. Cambridge University Press, 3 edition, 2014. ISBN 9781107038691. doi : 10.1017/CBO9781139839327.

E. L. Ionides, C. Breto, and A. A. King. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences of the United States of America, 103(49) :18438–18443, 2006. doi : 10.1073/pnas.0603181103.

E. L. Ionides, A. Bhadra, Y. Atchadé, and A. A. King. Iterated filtering. The Annals of Statistics, 39(3) :1776–1802, 2011. doi : 10.1214/11-aos886.

E. L. Ionides, D. Nguyen, Y. Atchadé, S. Stoev, and A. A. King. Inference for dynamic and latent variable models via iterated, perturbed bayes maps. Proceedings of the National Academy of Sciences of the United States of America, 112(3) :719–724, 2015. doi : 10.1073/pnas.1410597112.

E. L. Ionides, C. Breto, J. Park, R. A. Smith, and A. A. King. Monte carlo profile confidence intervals for dynamic systems. J. R. Soc Interface, 14 :2017126, 2017. doi : 10.1098/rsif.2017.0126.

A. M. Johnson, J. Wadsworth, K. Wellings, and J. Field. Sexual Attitudes and Lifestyles. Oxford, UK : Blackwell Scientific Publications, 1994.

M. J. Keeling and P. Rohani. Modeling Infectious Diseases in Humans and Animals. Princeton University Press, 2007.

W. O. Kermack, A. G. McKendrick, and G. T. Walker. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 115(772) :700–721, 1927. doi : 10.1098/rspa.1927.0118.

M. Kessler, A. Lindner, and M. Sorensen. Statistical Methods for Stochastic Differential Equations. Chapman and Hall/CRC, 2012. doi : 10.1201/b12126.

S. Khedhiri. Statistical modeling of covid-19 deaths with excess zero counts. Epidemiologic Methods, 10(s1) :20210007, 2021. doi : doi:10.1515/em-2021-0007.

C. T. Kiem, C. R. Massonnaud, D. Levy-Bruhl, C. Poletto, V. Colizza, P. Bosetti, A. Fontanet, A. Gabet, V. Olié, L. Zanetti, and et al. A modelling study investigating short and medium-term challenges for covid-19 vaccination : From prioritisation to the relaxation of measures. EClinicalMedicine, 38, 2021. doi : 10.1016/j.eclinm.2021.101001.

A. A. King, E. L. Ionides, M. Pascual, and M. J. Bouma. Inapparent infections and cholera dynamics. Nature, 454(7206) :877–880, 2008. doi : 10.1038/nature07084.

A. A. King, M. Domenech de Cellès, F. M. G. Magpantay, and P. Rohani. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to ebola. Proceedings of the Royal Society B : Biological Sciences, 282(1806) :20150347, 2015. doi : 10.1098/rspb.2015.0347.

A. A. King, D. Nguyen, and E. L. Ionides. Statistical inference for partially observed markov processes via the r package pomp. Journal of Statistical Software, 69(12) :1–43, 2017. doi : 10.18637/jss.v069.i12.

S. Kirkpatrick. Optimization by simulated annealing : Quantitative studies. J. Stat. Phys., 34 : 975–986, 1984. doi : 10.1007/BF01009452.

M. Kretzschmar, Y. T. van Duynhoven, and A. J. Severijnen. Modeling prevention strategies for gonorrhea and chlamydia using stochastic network simulations. American Journal of Epidemiology, 144(3) :306–317, 1996. doi : 10.1093/oxfordjournals.aje.a008926.

Y. Kuang, A. Tridane, and R. Lopez-Cruz. A simple si model with two age groups and its application to us hiv epidemics : To treat or not to treat ? Journal of Biological Systems, 2007.

E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of em with an mcmc procedure. ESAIM : Probability and Statistics, 8 :115–131, 2004. doi : 10.1051/ps:2004007.

E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. CSDA, 49(4) :1020–1038, 2005. doi : 10.1016/j.csda.2004.07.002.

M. Lavielle. Mixed Effects Models for the Population Approach : Models, Tasks, Methods and Tools (1st ed.). Chapman and Hall/CRC, 2014. doi : 10.1201/b17203.

J. S. Lavine, A. A. King, V. Andreasen, and O. N. Bjørnstad. Immune boosting explains regime-shifts in prevaccine-era pertussis dynamics. PLOS ONE, 8(8) :e72086, 2013. doi : 10.1371/journal.pone.0072086.

R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). Science, 368 (6490) :489–493, 2020. doi : 10.1126/science.abb3221.

Z. Liu, P. Magal, and G. Webb. Predicting the number of reported and unreported cases for the covid-19 epidemics in china, south korea, italy, france, germany and united kingdom. medRxiv, 2020. doi : 10.1101/2020.04.09.20058974v2.

A. L. Lloyd. Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. Proceedings of the Royal Society of London. Series B : Biological Sciences, 268(1470) :985–993, 2001. doi : 10.1098/rspb.2001.1599.

J-M Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate bayesian computational methods. Statistics and Computing, 22(6) :1167–1180, 2012. doi : 10.1007/s11222-011-9288-2.

C. J. E. Metcalf, O. N. Bjørnstad, B. T. Grenfell, and V. Andreasen. Seasonality and comparative dynamics of six childhood infections in pre-vaccination copenhagen. Proceedings. Biological Sciences, 276(1676) :4111–4118, 2009. doi : 10.1098/rspb.2009.1058.

C. E. Mills, J. M. Robins, and M. Lipsitch. Transmissibility of 1918 pandemic influenza. Nature, 432 :904–906, 2004. doi : 10.1038/nature03063.

R. Narci, M. Delattre, C. Larédo, and E. Vergu. Inference for partially observed epidemic dynamics guided by kalman filtering techniques. CSDA, 164, 2021a. doi : 10.1016/j.csda.2021.107319.

R. Narci, M. Delattre, C. Larédo, and E. Vergu. Inference in Gaussian state-space models with mixed effects for multiple epidemic dynamics. working paper or preprint, 2021b. URL https://hal.archives-ouvertes.fr/hal-03347365.

J. R. Norris. Markov chains. Cambridge University Press, 1997. doi : 10.1017/CBO9780511810633.

P. D. O'Neill. Introduction and snapshot review : Relating infectious disease transmission models to data. Statistics in Medicine, 29(20) :2069–2077, 2010. doi : 10.1002/sim.3968.

A. Pan, L. Liu, C. Wang, H. Guo, X. Hao, Q. Wang, J. Huang, N. He, H. Yu, X. Lin, and et al. Association of public health interventions with the epidemiology of the covid-19 outbreak in wuhan, china. JAMA, 323(19) :1915–1923, 2020. doi : 10.1001/jama.2020.6130.

S. W. Park and B. M. Bolker. A note on observation processes in epidemic models. Bulletin of Mathematical Biology, 82(3) :37, 2020. doi : 10.1007/s11538-020-00713-2.

Pasteur. Institut pasteur, fiche maladie de la grippe. https://www.pasteur.fr/fr/centre-medical/fiches-maladies/grippe#epidmiologie, 2020.

Pasteur. Institut pasteur, fiche maladie du virus ebola. https://www.pasteur.fr/fr/centre-medical/fiches-maladies/ebola#epidmiologie, 2021.

J. C. Pinheiro and D. M. Bates. Mixed-Effects Models in S and S-PLUS. Springer, 2000. doi : 10.1007/b98882.

B. Pourbohloul, A. Ahued, B. Davoudi, R. Meza, L. A. Meyers, D. M. Skowronski, I. Villaseñor, F. Galvan, P. Cravioto, D. J. Earn, J. Dushoff, D. Fisman, W. J. Edmunds, N. Huper, S. V. Scarpino, J. Trujillo, M. Lutzow, J. Morales, A. Contreras, C. Chavez, D. M. Patrick, and R. C. Brunham. Initial human transmission dynamics of the pandemic (h1n1) 2009 virus in north america. Influenza and Other Respiratory Viruses, 3(5) :215–222, 2009. doi : 10.1111/j.1750-2659.2009.00100.x.

M. Prague, L. Wittkop, A. Collin, D. Dutartre, Q. Clairon, P. Moireau, R. Thiébaut, and B. P. Hejblum. Multi-level modeling of early covid-19 epidemic dynamics in french regions and estimation of the lockdown impact on infection rate. medRxiv, 2020. doi : 10.1101/2020.04.21.20073536.

G. Pullano, L. Di Domenico, C. E. Sabbatini, E. Valdano, C. Turbelin, M. Debin, C. Guerrisi, C. Kengne-Kuetche, C. Souty, T. Hanslik, and et al. Underdetection of cases of covid-19 in france threatens epidemic control. Nature, 590(7844) :134–139, 2021. doi : 10.1038/s41586-020-03095-6.

Pejman Rohani, Matthew J. Keeling, and Bryan T. Grenfell. The interplay between determinism and stochasticity in childhood diseases. The American Naturalist, 159(5) :469–481, 2002. doi : 10.1086/339467.

J. V. Ross, D. E. Pagendam, and P. K. Polett. On parameter estimation in population models ii : Multidimensional processes and transient dynamics. Theoretical Population Biology, 75(2-3) : 123–132, 2009. doi : 10.1016/j.tpb.2008.12.002.

J. Roux, C. Massonnaud, V. Colizza, S. Cauchemez, and P. Crépey. Impact of national and regional lockdowns on covid-19 epidemic waves : Application to the 2020 spring wave in france. medRxiv, 2021. doi : 10.1101/2021.04.21.21255876v1.

J. Shaman and M. Kohn. Absolute humidity modulates influenza survival, transmission, and seasonality. Proceedings of the National Academy of Sciences, 106(9) :3243–3248, 2009. doi : 10.1073/pnas.0806852106.

S. Shrestha, A. A. King, and P. Rohani. Statistical inference for multi-pathogen systems. PLOS Computational Biology, 7(8) :e1002135, 2011. doi : 10.1371/journal.pcbi.1002135.

S. Shrestha, B. Foxman, D. M. Weinberger, C. Steiner, C. Viboud, and P. Rohani. Identifying the interaction between influenza and pneumococcal pneumonia using incidence data. Science Translational Medicine, 5(191) :191ra84, 2013. doi : 10.1126/scitranslmed.3005982.

S. Sisson, Y. Fan, and M. Tanaka. Sequential monte carlo without likelihoods. Proceedings of the National Academy of Sciences of the United States of America, 104(6) :1760–1765, 2007. doi : 10.1073/pnas.0607208104.

M. T. Sofonea, B. Reyné, B. Elie, R. Djidjou-Demasse, C. Selinger, Y. Michalakis, and S. Alizon. Epidemiological monitoring and control perspectives : application of a parsimonious modelling framework to the covid-19 dynamics in france. medRxiv, 2020. doi : 10.1101/2020.05.22.20110593.

T. Stocks. Iterated filtering methods for markov process epidemic models, 2017.

T. Stocks, T. Britton, and M. Höhle. Model selection and parameter estimation for dynamic epidemic models via iterated filtering : application to rotavirus in germany. Biostatistics, 21(3) : 400–416, 2018. doi : 10.1093/biostatistics/kxy057.

S. B. Stoecklin, P. Rolland, Y. Silue, A. Mailles, C. Campese, A. Simondon, M. Mechain, L. Meurice, M. Nguyen, C. Bassi, and et al. First cases of coronavirus disease 2019 (covid-19) in france : surveillance, investigations and control measures, january 2020. Eurosurveillance, 25 (6) :2000094, 2020. doi : 10.2807/1560-7917.ES.2020.25.6.2000094.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P.H Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of The Royal Society Interface, 6(31) :187–202, 2009. doi : 10.1098/rsif.2008.0172.

P. van den Driessche and J. Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Mathematical Biosciences, 180(1) : 29–48, 2002. doi : 10.1016/S0025-5564(02)00108-6.

G. C. G. Wei and M. A. Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. JASA, 85 :699–704, 1990.

WHO. Who coronavirus disease (covid-19) dashboard. https://covid19.who.int., Octobre 2021.

C. F. Jeff Wu. On the convergence properties of the em algorithm. The Annals of Statistics, 11 (1) :95–103, 1983. doi : 10.1214/aos/1176346060.